

VIETNAM SCHOLARHUB — HƯỚNG DẪN PHƯƠNG PHÁP NGHIÊN CỨU

Tính toán Cỡ mẫu & Power Analysis trong Nghiên cứu Khoa học

*Nền tảng không thể bỏ qua trước khi thiết kế bất kỳ nghiên cứu nào***70–90%**

bài bị desk reject

80%

power tối thiểu cần đạt

0.05

ngưỡng alpha tiêu chuẩn

6

yếu tố ảnh hưởng cỡ mẫu

01 Lý do thiếu mẫu là nguyên nhân thâm lạng của hàng ngàn bài báo bị từ chối

Trong thực tiễn biên tập tại các tạp chí khoa học quốc tế, một trong những lý do phổ biến nhất khiến bản thảo bị từ chối ngay từ vòng desk review không phải vì câu hỏi nghiên cứu thiếu giá trị, cũng không phải vì dữ liệu không đầy đủ — mà vì **cỡ mẫu không đủ để hỗ trợ cho kết luận được đưa ra**. Reviewer có kinh nghiệm nhìn vào phần Methods là nhận ra ngay liệu tác giả có thực sự kiểm soát được độ tin cậy thống kê hay không.

Chuỗi nhân quả cần ghi nhớ:

Không đủ mẫu → Không đủ statistical power → Kết quả không thuyết phục reviewer
→ Desk rejection hoặc major revision không thể vượt qua

Điều đáng nói là phần lớn các nhà nghiên cứu lâm sàng — kể cả những người có nhiều năm kinh nghiệm điều trị — thường hiểu sai hoặc chưa hiểu đầy đủ về ý nghĩa của power analysis. Nhiều người vẫn xem đây là thủ tục hành chính chỉ để điền vào biểu mẫu đề cương. Trong khi thực tế, **power analysis là điểm khởi đầu bắt buộc của toàn bộ thiết kế nghiên cứu**, không phải bước bổ sung ở cuối quy trình.

Một trong những bằng chứng nổi bật nhất đến từ phân tích hệ thống của Button và cộng sự (2013) đăng trên Nature Reviews Neuroscience, chỉ ra rằng median statistical power trong các

nghiên cứu thần kinh học chỉ đạt khoảng **8–31%** — tức là phần lớn nghiên cứu được thiết kế với power quá thấp để phát hiện các hiệu ứng quan trọng. Tình trạng tương tự được ghi nhận trong y học lâm sàng, tâm lý học và nhiều ngành khoa học sức khỏe khác.

02 Power Analysis là gì? — Nguyên lý vận hành và ý nghĩa thực tiễn

Power analysis — hay phân tích lực thống kê — là quá trình xác định cỡ mẫu tối thiểu cần thiết để một nghiên cứu có khả năng phát hiện được sự khác biệt thực sự (effect size) trong quần thể, với một mức độ tin cậy và lực thống kê được xác định trước.

Hai loại sai lầm thống kê mà mọi nhà nghiên cứu phải nắm

✗ Type I Error (Sai lầm loại I)	✓ Type II Error (Sai lầm loại II)
Kết luận có sự khác biệt khi thực tế KHÔNG có	Kết luận KHÔNG có sự khác biệt khi thực tế có
Dương tính giả (False Positive)	Âm tính giả (False Negative)
Kiểm soát bằng ngưỡng alpha (thường 0.05)	Kiểm soát bằng power (thường 80%)
Xảy ra khi alpha đặt quá cao hoặc phân tích nhiều biến không điều chỉnh	Xảy ra khi cỡ mẫu quá nhỏ — phổ biến hơn nhiều trong thực tế
Hậu quả: kết quả dương tính giả được công bố và khó thu hồi	Hậu quả: bỏ lỡ phát hiện quan trọng, nghiên cứu lãng phí nguồn lực

Bẫy nhận thức phổ biến nhất:

Kết quả không có ý nghĩa thống kê ($p > 0.05$) **KHÔNG đồng nghĩa** với việc không có sự khác biệt. Rất có thể nghiên cứu không đủ power để nhìn thấy sự khác biệt đó. Reviewer tại các tạp chí Q1 sẽ luôn hỏi: "Nếu kết quả âm tính, power của nghiên cứu là bao nhiêu?". Nếu tác giả không có câu trả lời, bài báo gần như chắc chắn bị reject hoặc yêu cầu major revision toàn diện.

Một hình ảnh trực quan: power giống như độ phân giải của kính hiển vi. Kính hiển vi có độ phân giải thấp sẽ không thể quan sát được vi khuẩn dù vi khuẩn đang hiện diện rõ ràng trong mẫu. Tương tự, nghiên cứu với power thấp không thể phát hiện hiệu ứng dù hiệu ứng đó thực sự tồn tại trong quần thể — và kết quả âm tính như vậy không mang nhiều giá trị khoa học.

03 Sáu yếu tố quyết định cỡ mẫu — Nguyên lý vận hành đằng sau con số

Thay vì tiếp cận theo hướng công thức, phần này phân tích nguyên lý vận hành đằng sau từng yếu tố, giúp nhà nghiên cứu hiểu tại sao các quyết định thiết kế lại ảnh hưởng trực tiếp đến quy mô mẫu — và từ đó đưa ra quyết định thiết kế thông minh hơn ngay từ giai đoạn lên kế hoạch.

Yếu tố	Tác động đến cỡ mẫu	Ví dụ thực tiễn
1. Thiết kế nghiên cứu	Yếu tố có ảnh hưởng lớn nhất, có thể giảm hoặc tăng mẫu từ 20–75%	<i>Cross-over cần ¼ mẫu so với parallel; pre-post cần ½ so với có nhóm chứng riêng</i>
2. Loại biến kết cục	Biến liên tục cần ít mẫu hơn biến nhị phân vì độ nhạy thống kê cao hơn	<i>Huyết áp (liên tục) vs. có/không mắc bệnh (nhị phân)</i>
3. Alpha (α)	Alpha càng thấp → cần mẫu càng lớn để đảm bảo kiểm soát Type I error	<i>$\alpha=0.05$ cho nghiên cứu thông thường; $\alpha=0.01$ cho nghiên cứu ảnh hưởng chính sách lớn</i>
4. Độ lệch chuẩn (SD)	Biến động càng lớn → tín hiệu càng khó phân biệt → cần nhiều mẫu hơn	<i>TNF-alpha (biến động lớn) cần nhiều mẫu hơn cân nặng sơ sinh</i>
5. Effect size (MDD)	Muốn phát hiện khác biệt nhỏ → cần mẫu rất lớn; khác biệt lớn → cần ít mẫu	<i>Giảm 2 mmHg huyết áp cần mẫu lớn hơn nhiều so với giảm 15 mmHg</i>
6. Power (1- β)	Power 80% là tối thiểu; 90% cho nghiên cứu quan trọng về chính sách/điều trị	<i>Power 90% cần tăng ~30% cỡ mẫu so với power 80%</i>

3.1 Thiết kế nghiên cứu — Yếu tố có đòn bẩy mạnh nhất

Thiết kế nghiên cứu là yếu tố đầu tiên và có tác động lớn nhất đến cỡ mẫu. Một quyết định thiết kế thông minh có thể giảm đến 75% số lượng người cần tham gia, trong khi một thiết kế không phù hợp sẽ đẩy quy mô mẫu lên mức không khả thi về nguồn lực.

Nguyên tắc điều chỉnh cỡ mẫu theo thiết kế:

- Thiết kế bắt chéo (cross-over): cần khoảng 1/4 mẫu so với thiết kế song song có nhóm chứng riêng biệt — vì mỗi người trải qua cả hai điều kiện, loại trừ biến thể giữa các cá thể (between-subject variability)
- Thiết kế tiền-sau (pre-post) một nhóm: cần khoảng 1/2 mẫu so với thiết kế có nhóm đối chứng tương đương, do mỗi người tự làm chứng cho chính mình
- Giả thuyết một chiều (one-tailed): giảm được khoảng 20% cỡ mẫu so với hai chiều — nhưng chỉ sử dụng khi có cơ sở lý thuyết rõ ràng về hướng tác động
- Nghiên cứu không thể randomize: cộng thêm tối thiểu 20% để bù đắp cho nhiễu do yếu tố confounding không kiểm soát được
- Dự phòng attrition: luôn cộng thêm 10–20% cho các trường hợp bỏ cuộc, mất dấu theo dõi (loss to follow-up) hoặc dữ liệu bị thiếu (missing data)

3.2 Biến kết cục — Lựa chọn chiến lược ảnh hưởng đến tính khả thi

Loại biến kết cục ảnh hưởng đến cỡ mẫu theo cơ chế trực tiếp: các phép kiểm định thống kê áp dụng cho biến liên tục (như t-test hay ANOVA) có độ nhạy cao hơn so với các phép kiểm định cho biến nhị phân (như chi-square hay Fisher's exact). Do đó, cùng một effect size thực sự, nghiên cứu sử dụng biến liên tục sẽ cần ít người hơn để phát hiện.

✗ Biến cần mẫu lớn hơn	✓ Biến cần mẫu nhỏ hơn
Biến nhị phân: có/không (tiêm vaccine, mắc bệnh)	Biến liên tục: huyết áp, cân nặng, nồng độ IL-6, VAS
Biến danh nghĩa nhiều mức loại trừ: nhóm máu, tín ngưỡng	Biến thứ tự với nhiều mức: thang điểm chức năng, chất lượng sống
Cần mẫu lớn hơn đáng kể	Cần mẫu nhỏ hơn với cùng effect size

Khó kiểm soát confounding bằng thống kê

Linh hoạt hơn trong phân tích hồi quy và điều chỉnh covariates

Lưu ý quan trọng: Việc lựa chọn biến kết cục không chỉ ảnh hưởng đến cỡ mẫu mà còn quyết định ý nghĩa lâm sàng và tính khả thi của nghiên cứu. Một biến kết cục được đo lường chính xác và có liên quan trực tiếp đến câu hỏi nghiên cứu luôn được ưu tiên hơn biến thay thế (surrogate endpoint) chỉ vì lý do tiết kiệm mẫu.

3.3 Effect Size và Minimum Detectable Difference — Thông số nhạy cảm nhất

Effect size — hay còn gọi là minimum detectable difference (MDD) — là yếu tố nhạy cảm nhất trong tính toán cỡ mẫu. Nó đại diện cho sự khác biệt nhỏ nhất có ý nghĩa lâm sàng mà nghiên cứu cần có khả năng phát hiện. Mối quan hệ giữa effect size và cỡ mẫu là quan hệ nghịch biến phi tuyến: muốn phát hiện sự khác biệt nhỏ hơn, cỡ mẫu cần tăng theo hàm số mũ, không phải tuyến tính.

Câu hỏi thực tiễn mà nhiều nhà nghiên cứu bế tắc:

"Làm sao biết trước effect size khi chưa thực hiện nghiên cứu?"

Có ba nguồn tham khảo hợp lệ và được reviewer chấp nhận:

- Tổng quan hệ thống và meta-analysis trong lĩnh vực liên quan — nguồn đáng tin cậy nhất
- Nghiên cứu pilot hoặc dữ liệu nền (baseline data) từ cùng bệnh viện hoặc cộng đồng
- Phán đoán lâm sàng có căn cứ: nhà nghiên cứu lâm sàng xác định ngưỡng khác biệt nào là đủ quan trọng để thay đổi quyết định điều trị

Lưu ý: Cần phân biệt rõ giữa **statistical significance** và **clinical significance**. Một nghiên cứu với cỡ mẫu rất lớn có thể phát hiện sự khác biệt 0.3 mmHg huyết áp với $p < 0.001$ — nhưng con số đó không có bất kỳ ý nghĩa điều trị nào. Chọn effect size phải dựa trên ngưỡng lâm sàng có ý nghĩa, không phải bất kỳ con số nào cho ra $p < 0.05$.

3.4 Alpha, Power và Độ lệch chuẩn — Ba thông số còn lại

Alpha (α) kiểm soát xác suất mắc Type I error. Tiêu chuẩn phổ biến nhất là $\alpha = 0.05$, nghĩa là chấp nhận 5% khả năng kết luận sai. Với các nghiên cứu có ảnh hưởng lớn đến chính sách quốc gia hoặc quyết định điều trị phổ biến, tiêu chuẩn nghiêm ngặt hơn như $\alpha = 0.01$ hoặc $\alpha = 0.005$ (ngưỡng được đề xuất bởi Benjamin và cộng sự, 2018 trên Nature Human Behaviour) đang ngày càng được áp dụng rộng rãi.

Power ($1-\beta$) — thường được đặt ở mức 80% — là xác suất để phát hiện được sự khác biệt nếu nó thực sự tồn tại. Đây là xác suất tránh được Type II error. Power 90% được khuyến cáo cho các nghiên cứu thử nghiệm lâm sàng pha III, nghiên cứu có tác động chính sách lớn, hoặc khi kết quả âm tính cũng có giá trị công bố quan trọng. Tăng từ 80% lên 90% thường đòi hỏi tăng cỡ mẫu thêm khoảng 30%.

Độ lệch chuẩn (SD) của biến kết cục phản ánh mức độ biến thiên tự nhiên của dữ liệu trong quần thể nghiên cứu. Biến thiên càng lớn, tín hiệu từ can thiệp càng bị "chìm" trong nhiễu nền, và do đó cần nhiều mẫu hơn để phân biệt được. Giá trị SD nên được lấy từ nghiên cứu trước trên quần thể tương đồng hoặc từ pilot study — không nên ước lượng chủ quan.

04 Power Analysis trước và sau thu thập số liệu — Hai ứng dụng quan trọng

Power analysis không chỉ là công cụ lên kế hoạch — nó còn là công cụ diễn giải kết quả sau khi nghiên cứu hoàn tất. Cần phân biệt rõ hai loại power analysis này, vì chúng có mục đích và cách báo cáo khác nhau hoàn toàn.

A priori (Lập kế hoạch)	Post hoc (Diễn giải)
A priori power analysis — Trước thu thập số liệu	Post hoc power analysis — Sau phân tích dữ liệu
Xác định cỡ mẫu tối thiểu cần thiết	Giải thích kết quả âm tính ($p > 0.05$)
Input: alpha, power mong muốn, effect size ước lượng, SD	Input: cỡ mẫu thực tế, alpha, effect size quan sát được
Output: N cần thiết → lên kế hoạch tuyển chọn đối tượng	Output: power thực tế → đánh giá mức độ kết luận được từ kết quả

Bắt buộc trong hầu hết journal Q1/Q2 hiện nay	Quan trọng khi báo cáo kết quả âm tính; cần giải thích rõ
Nên thực hiện trước khi đăng ký đề cương / pre-registration	Nếu power < 50%: kết quả âm tính không thể kết luận gì đáng tin

Xu hướng mới tại các tạp chí Q1 (2022–2024):

- Nhiều tạp chí như JAMA, Lancet, BMJ hiện yêu cầu pre-registration kèm power analysis trước khi thu thập số liệu (clinicaltrials.gov hoặc OSF)
- Một số journal áp dụng Registered Reports — chấp nhận bài dựa trên đề cương và power analysis, trước khi biết kết quả — nhằm chống publication bias
- Báo cáo kết quả âm tính (null results) được khuyến khích hơn nếu power $\geq 80\%$ — vì kết quả âm tính có power cao cũng là bằng chứng khoa học giá trị

05 Phân biệt ý nghĩa thống kê và ý nghĩa lâm sàng — Ranh giới quan trọng nhất

Một điểm then chốt mà nhiều nhà nghiên cứu — đặc biệt những người mới bắt đầu công bố quốc tế — thường nhầm lẫn: **$p < 0.05$ không đồng nghĩa với ý nghĩa lâm sàng**. Đây là hai khái niệm hoàn toàn độc lập, và reviewer tại các tạp chí hàng đầu sẽ luôn đặt câu hỏi về cả hai.

✗ Tiếp cận chỉ theo p-value	✓ Tiếp cận kết hợp thống kê và lâm sàng
Chỉ cần $p < 0.05$ là đủ để kết luận	Cần cả $p < 0.05$ VÀ effect size có ý nghĩa lâm sàng
Cỡ mẫu rất lớn $\rightarrow$ phát hiện mọi khác biệt dù rất nhỏ	Xác định trước ngưỡng khác biệt có giá trị điều trị
Giảm huyết áp 0.5 mmHg: $p < 0.001$ với $n=50.000$	Giảm huyết áp 5 mmHg: $p < 0.05$ với $n=200$ — có ý nghĩa thực hành
Dễ bị "p-hacking" hoặc HARKing	Đòi hỏi xác định outcome và effect size trước khi thu số liệu

Nguy cơ: kết quả không thể tái lập, gây hiểu lầm chính sách

Kết quả robust, có giá trị ứng dụng thực tiễn

Hướng dẫn thực hành:

Trước khi bắt đầu nghiên cứu, nhóm tác giả cần xác định rõ và ghi vào đề cương:

1. Đây là mức khác biệt tối thiểu có ý nghĩa lâm sàng (MCID — Minimum Clinically Important Difference)?
2. Effect size này có được hỗ trợ bởi y văn hiện tại không?
3. Nghiên cứu được thiết kế để phát hiện effect size đó với power bao nhiêu?

Đây là ba câu hỏi mà reviewer tại NEJM, JAMA, Lancet sẽ luôn đặt ra. Trả lời được ba câu này trong phần Methods chính là bằng chứng cho thấy đây là một nghiên cứu được thiết kế nghiêm túc và minh bạch.

06 Checklist thực hành — Những gì cần làm trước khi bắt đầu bất kỳ nghiên cứu nào

Dưới đây là checklist được biên soạn dựa trên hướng dẫn của CONSORT 2010 (cho RCT), STROBE (cho nghiên cứu quan sát), và PRISMA (cho tổng quan hệ thống) — ba bộ tiêu chuẩn báo cáo phổ biến nhất trong y văn quốc tế.

Checklist Power Analysis & Sample Size:

- ✓ Xác định rõ biến kết cục chính (primary outcome) — phân biệt với các biến phụ (secondary outcomes)
- ✓ Xác định loại biến kết cục: nhị phân, danh nghĩa, thứ tự, liên tục — để chọn phép kiểm định và ước lượng mẫu phù hợp
- ✓ Tra cứu y văn để xác định effect size kỳ vọng (MDD/MCID) có căn cứ khoa học
- ✓ Xác định SD của biến kết cục từ nghiên cứu trước hoặc pilot study — không ước đoán chủ quan
- ✓ Đặt ngưỡng alpha phù hợp với mục tiêu nghiên cứu (0.05 cho hầu hết; 0.01 cho nghiên cứu ảnh hưởng chính sách lớn)
- ✓ Đặt power tối thiểu 80% (90% cho RCT pha III hoặc nghiên cứu có giá trị chính sách cao)

- ✓ Tính cỡ mẫu bằng phần mềm chuyên dụng: G*Power (miễn phí), PASS, hoặc R packages (pwr, WebPower)
- ✓ Cộng thêm 10–20% dự phòng cho attrition và missing data — nêu rõ tỷ lệ giả định trong đề cương
- ✓ Báo cáo đầy đủ trong phần Methods: phần mềm sử dụng, các thông số đầu vào, kết quả đầu ra — không cần viết công thức
- ✓ Xem xét pre-registration tại ClinicalTrials.gov, ANZCTR, hoặc OSF trước khi bắt đầu thu thập số liệu

Lưu ý về cách báo cáo trong bài báo

Power analysis trong phần Methods của bài báo **không cần và không nên** viết ra công thức toán học. Cách tiếp cận này vừa không cần thiết về mặt khoa học, vừa làm nặng nề phần Methods và giảm tính đọc được của bài báo.

Ví dụ cách báo cáo chuẩn (được reviewer Q1 chấp nhận):

"Sample size was calculated using G*Power 3.1. Assuming a mean difference of 5 mmHg in systolic blood pressure (SD = 12 mmHg, based on [reference]), with a two-sided alpha of 0.05 and 80% power, a minimum of 91 participants per group was required. Accounting for an estimated 15% dropout rate, we planned to enroll 108 participants per group."

Đây là tất cả thông tin mà reviewer cần để đánh giá tính hợp lý của cỡ mẫu. Không cần gì thêm.

07 Kết luận — Power Analysis là tư duy nghiên cứu, không phải thủ tục

Power analysis không phải là bước cuối cùng được thêm vào đề cương để đủ biểu mẫu. Đó là **bằng chứng của tư duy nghiên cứu có hệ thống**: biết mình đang tìm gì, cần bao nhiêu bằng chứng để kết luận, và sẵn sàng đối mặt với kết quả âm tính có giá trị như kết quả dương tính.

Làm nghiên cứu mà không có power analysis được thiết kế kỹ lưỡng giống như phẫu thuật viên bước vào phòng mổ mà không biết rõ mục tiêu phẫu thuật, không kiểm tra đủ

dụng cụ, và chỉ hi vọng may mắn mỉm cười. Cả hai trường hợp đều có thể dẫn đến hậu quả nghiêm trọng — và cả hai đều có thể phòng tránh được bằng sự chuẩn bị kỹ lưỡng từ trước.

Với bất kỳ nhà nghiên cứu nào đang lên kế hoạch cho một đề tài — dù là luận văn thạc sĩ, đề tài cấp cơ sở hay thử nghiệm lâm sàng ngẫu nhiên — bước đầu tiên nên là ngồi lại với câu hỏi: **"Nghiên cứu này cần bao nhiêu người để kết quả của nó thực sự đáng tin?"** Câu trả lời cho câu hỏi đó chính là xuất phát điểm của một nghiên cứu nghiêm túc.