Power Analysis & Sample Size Determination

A Quick Reference for Medical Researchers

Core Concepts

A **power analysis** answers a practical question:

> *"How many patients do I need to enroll to have a good chance of detecting a clinically meaningful difference?"*

Every sample size calculation is built around four linked pieces:

Concept	Meaning	Typical Value
Significance level (α)	Probability of a false positive (finding a difference that is not truly there)	0.05
Power ($1 - \beta$)	Probability of detecting a real effect if one truly exists	0.80 (80%)
Effect size	How big of a difference you expect between groups	Based on prior studies
Sample size (n)	Number of subjects needed	Usually what we solve for

The key tradeoffs:

- Smaller expected differences → require **larger sample sizes**
- Higher desired power → requires **larger sample sizes**
- More variability ("noise") in the data → requires **larger sample sizes**

A Common Misconception

A frequent question is *"How many patients do I need to get a p-value < 0.05?"* or *"How many patients do I need for statistical significance?"* This is **not** what power analysis answers, and framing it that way is misleading in two important ways:

1. **You cannot design a study to guarantee statistical significance.** A p-value is a result that depends on the data you collect, not a target you can set. Whether your study produces $p < 0.05$ depends on whether a real effect exists and how large it actually is — neither of which you control by enrolling more patients.
2. **If no true effect exists, no sample size will reliably produce a "significant" result.** Larger samples only help you detect effects that are actually there. Chasing significance with ever-larger enrollment, repeated looks at the data, or multiple analyses is a hallmark of **p-hacking**, not good study design.

What power analysis *actually* answers is:

"Assuming a clinically meaningful effect of a certain size truly exists, how many patients do I need to have a high probability of detecting it?"

What is an Effect Size?

An **effect size** is a standardized measure of how big a difference is, expressed in a way that does not depend on the units of measurement or the size of the study. It is the single most important assumption in any power analysis — and the one most commonly misunderstood.

Why we standardize

A raw difference — "5 mmHg" or "10% fewer readmissions" — is meaningful clinically, but on its own it doesn't tell statistical software how *detectable* that difference is. **Detectability depends on how the difference compares to the underlying variability in the data.**

Consider two scenarios where the BP difference between groups is the same 5 mmHg:

- **Scenario A:** Standard deviation = 2 mmHg. The 5 mmHg gap is huge relative to noise — easy to detect, small sample needed.
- **Scenario B:** Standard deviation = 50 mmHg. The 5 mmHg gap is buried in noise — very hard to detect, large sample needed.

The same 5 mmHg difference can require dramatically different sample sizes depending on the noise. Effect size combines the difference *and* the variability into one number so the same statistical tools work regardless of what you are measuring.

Think of it as signal-to-noise

A useful mental model:

$$\text{Effect size} \approx \frac{\text{signal (group difference)}}{\text{noise (variability)}}$$

- Bigger signal → larger effect size → easier to detect → smaller sample needed
- Bigger noise → smaller effect size → harder to detect → larger sample needed

This is why two studies looking at the "same" 5 mmHg difference can need wildly different enrollments: it is the *standardized* effect, not the raw difference, that drives sample size.

Effect size measures by outcome type

For **continuous outcomes** (like blood pressure or HbA1c), the standard measure is **Cohen's** d — the difference in means divided by the standard deviation. This is what the `pwr` package uses for t-test calculations.

For **binary outcomes** (proportions, like cancer or readmission rates), there is usually no need to convert to a standardized effect size. Rates are already on a common 0–1 scale, so most power tools take the two proportions directly. This keeps the calculation close to how clinicians actually think about the comparison ("we expect 30% to drop to 15%") rather than forcing an extra layer of arithmetic.

A warning about Cohen's benchmarks

The conventional labels "small / medium / large" (e.g., $d = 0.2$ / 0.5 / 0.8) are **statistical conventions, not clinical judgments**. A "small" effect of $d = 0.2$ might be enormously important clinically — a modest mortality reduction in a common disease saves many lives. A "large" effect of $d = 0.8$ might be clinically trivial in a setting where only very large changes matter. Always anchor effect sizes in **what is clinically meaningful for your patients**, not in Cohen's labels.

Where Do Effect Sizes Come From?

The most important assumption in a power analysis is the **effect size**. This should come from:

- Prior literature
- Pilot data
- Clinically meaningful differences

The previous study does **not** need to be identical to yours. Look for studies with similar patient populations, similar outcomes, and similar interventions.

Examples of what to extract

- Mean HbA1c reduction and standard deviation
- Mortality rates
- Readmission rates

If published estimates differ, it is usually safer to use the **smaller / more conservative effect size**, because overestimating the treatment effect can leave your study **underpowered** (too small to detect the true effect).

Example 1: Continuous Outcome (t-test)

Suppose prior studies show:

- Standard care lowers systolic BP by ~6 mmHg
- Your intervention is expected to lower it by ~11 mmHg
- Standard deviation is ~12 mmHg

So the expected difference between groups is:

$$11 - 6 = 5 \text{ mmHg}$$

Cohen's d standardizes this difference:

$$d = \frac{5}{12} \approx 0.42$$

Interpretation of Cohen's d:

d	Interpretation
0.2	Small effect
0.5	Medium effect
0.8	Large effect

```r
# Two-arm t-test (n = subjects per group)
pwr.t.test(
  d         = 0.42,
  sig.level = 0.05,
  power     = 0.80,
  type      = "two.sample"
)
#>
#>      Two-sample t test power calculation
#>
#>               n = 89.95986
#>               d = 0.42
#>       sig.level = 0.05
#>           power = 0.8
#>     alternative = two.sided
#>
#> NOTE: n is number in *each* group
```

Reading the output

The function returns `n = 89.96`. The note `n is number in *each* group` is critical: this means you need ~**90 patients per group**, or roughly **180 patients total**. Sample sizes are *always rounded up* (you cannot enroll a fraction of a patient, and rounding down would leave the study slightly underpowered).

Example 2: Binary Outcome (Two Proportions)

Suppose you are studying a skin lesion that, if left untreated, develops into cancer in **30% of patients**. An existing drug already reduces this rate to 20%. Your team is developing a new drug, and the company has decided it will only be worthwhile to pursue if it can bring the cancer rate down to **15% or better**.

So the comparison you want to power for is:

- Control (untreated) cancer rate: **30%**
- Expected new-drug cancer rate: **15%**

Clinical interpretation:

- Absolute risk reduction = 15 percentage points
- Relative risk reduction = 50%

For two-proportion comparisons, R's built-in `power.prop.test()` takes the two proportions directly — no need to convert to a standardized effect size first:

```r
power.prop.test(
  p1         = 0.30,
  p2         = 0.15,
  sig.level = 0.05,
  power      = 0.80
)
#>
#>      Two-sample comparison of proportions power calculation
#>
#>               n = 120.4719
#>              p1 = 0.3
#>              p2 = 0.15
#>       sig.level = 0.05
#>           power = 0.8
#>     alternative = two.sided
#>
#> NOTE: n is number in *each* group
```

Reading the output

The function reports `n` per group (round up any fractional value — you cannot enroll part of a patient). For this comparison you need roughly **121 patients per group**, or about **242 patients total**, assuming equal allocation between arms.

Interpreting Sample Size Results

The output from a power-analysis function usually gives:

- Required sample size per group
- Assumed effect size (or the input proportions, for binary outcomes)
- Alpha
- Power

Example interpretation:

> "We estimate that 90 patients per group (180 total) are needed to detect a 5 mmHg difference in systolic BP with 80% power at $\alpha = 0.05$, assuming a standard deviation of 12 mmHg (Cohen's $d \approx 0.42$)."

Always state:

1. The primary outcome
2. The expected effect size (or the assumed event rates, for binary outcomes)
3. The source of the assumptions
4. The planned power and alpha

Practical Tips

- Increase your planned sample size by ~10–20% to account for dropout, loss to follow-up, and missing data. (E.g., 90 per group becomes ~100–108 per group after a 15% dropout buffer.)
- Avoid designing studies around "best-case" effects — smaller, realistic effects are safer assumptions.
- A statistically significant result is not always clinically important, and the goal of a power analysis is not to "guarantee" $p < 0.05$. Power analysis should focus on a **clinically meaningful difference** — if such a difference truly exists, you want a strong chance of detecting it.
- Underpowered studies are a major reason studies fail to show significance even when treatments may truly work.
- For more advanced designs (survival analysis, repeated measures, non-inferiority, cluster trials), involve a biostatistician early.

One-Sentence Takeaway

> Power analysis is the process of estimating how many patients are needed to reliably detect a clinically meaningful effect while minimizing the risk of false-negative results.