

# e<sup>ecp</sup>

EFFECTIVE CLINICAL PRACTICE

## COMPENDIUM OF PRIMERS

### **Evaluating Statistics**

---

Probability and Odds and Interpreting Their Ratios

---

Type I and Type II Errors

---

Absolute vs. Relative Differences

---

Correlation Coefficients

---

95% CIs for the Number Needed to Treat

---

Statistical Significance and P values

---

95% Confidence Intervals

---

### **Evaluating Study Designs**

---

Before–After Studies: Evaluating a Report of a “Successful” Intervention

---

Group Randomized Trials

---

Cost-Effectiveness Analysis

---

Interpreting Surveys

---

Utilities

---

Scores: What Counts?

---

### **Special Topics**

---

Lead-Time, Length, and Overdiagnosis Biases

---

Dissecting a Medical Imperative

---

HEDIS

---

Geographic Variation in Health Care

---

## Primer on Probability and Odds and Interpreting Their Ratios

Chance is measured by using either probabilities (a ratio of occurrence to the whole) or odds (a ratio of occurrence to non-occurrence). Consider measuring the chance of breast-feeding among 1000 new mothers. If 600 ultimately breast-feed, the prob-

ability of breast-feeding is 600/1000, or 0.6 (often expressed as 60%), whereas the odds of breast-feeding are 600/400, or 1.5 (often expressed as 1.5 to 1). Table 1 summarizes the characteristics of probability and odds.

TABLE 1

CHARACTERISTIC	PROBABILITY	ODDS
Ratio	$\frac{\text{occurrence}}{\text{whole}}$	$\frac{\text{occurrence}}{\text{nonoccurrence}}$
Range	0 to 1	0 to $\infty$
Transformation to other measure	$\text{odds} = \frac{\text{probability}}{1 - \text{probability}}$	$\text{probability} = \frac{\text{odds}}{1 + \text{odds}}$

Probabilities and odds contain the same information and are equally valid measures of chance. In the case of infrequent events (i.e., probability < 0.1 or 10%), the distinction is unimportant (probability and odds have essentially the same value). However, as shown in Table 2, probability and odds take on very different values as the chance of an event increases.

Although probabilities are often reported in the medical literature, it is rare to see odds reported. On the other hand, ratios of probabilities (i.e., relative risks, or risk ratios [RRs]) and odds (i.e., odds ratios [ORs]) are seen often. And it is in these ratios of ratios that the distinction between probability and odds may be both important and ambiguous.

When the chance of common events are being compared, ORs and RRs substantially diverge in value. Let's return to the breast-feeding example. Imagine a randomized trial of a lactation-support system. The probability of breast-feeding in the control group is 60% (or an odds of 1.5); in the intervention group, it is 90% (or an odds of 9). Table 3 shows that the relative risk is 1.5 while the odds ratio is 6.

TABLE 2

PROBABILITY	ODDS
1	$\infty$
0.9	9.00
0.8	4.00
0.7	2.33
0.6	1.50
0.5	1.00
0.4	0.67
0.3	0.43
0.2	0.25
0.1	0.11
0	0.00

TABLE 3

GROUP	PROBABILITY OF BREAST-FEEDING	ODDS OF BREAST-FEEDING	RELATIVE RISK (INTERVENTION VS. CONTROL)	ODDS RATIO (INTERVENTION VS. CONTROL)
Control	0.6	1.5	0.9/0.6	9.0/1.5
Intervention	0.9	9.0	=1.5	=6

In general, ORs are more extreme (i.e., farther away from 1) than are RRs. ORs that are greater than 1 exaggerate the increase in risk (i.e., OR > RR); ORs that are less than 1 exaggerate the decrease in risk (i.e., OR < RR). Practically speaking, the discrepancy between the two measures is relevant only when relatively common events are being compared. Readers should begin to worry about the distinction when baseline probabilities exceed 10% to 20%. And, as shown in Table 4, they might reasonably pursue a conversion when baseline probabilities are greater than 50%.

It is important to emphasize that ORs and RRs are equally valid—but different—measures. Readers are seeing more and

more ORs in the medical literature, largely because of the increased use of logistic regression. Because most people are more familiar with probabilities than odds, ORs are often interpreted as RRs. When events are common, this misinterpretation substantially exaggerates the association being reported. If the goal is clarity, the probability (or absolute event rate) for each group is tough to beat.

**Suggested Reading**

Talryn H, Davies O, Crombie IK, Tavakoli M. When can odds ratios mislead? *BMJ*. 1998;316:989-91.

Zhang J, Yu KF. What's the relative risk? A method of correcting the odds ratios in cohort studies of common outcomes. *JAMA*. 1998;280:1690-1.

**TABLE 4**

APPROXIMATE RELATIVE RISK FOR ODDS RATIOS GREATER THAN 1						APPROXIMATE RELATIVE RISK FOR ODDS RATIOS LESS THAN 1					
ODDS RATIO	BASELINE PROBABILITY					ODDS RATIO	BASELINE PROBABILITY				
	50%	60%	70%	80%	90%		50%	60%	70%	80%	90%
1.2	1.09	1.07	1.05	1.03	1.02	0.9	0.95	0.96	0.97	0.98	0.99
1.5	1.20	1.15	1.11	1.07	1.03	0.8	0.89	0.91	0.93	0.95	0.98
2	1.33	1.25	1.18	1.11	1.05	0.5	0.67	0.71	0.77	0.83	0.91
5	1.67	1.47	1.32	1.19	1.09	0.3	0.46	0.52	0.59	0.68	0.81
10	1.82	1.56	1.37	1.22	1.10	0.1	0.18	0.22	0.27	0.36	0.53

## Primer on Type I and Type II Errors

Statistical tests are tools that help us assess the role of chance as an explanation of patterns observed in data. The most common “pattern” of interest is how two groups compare in terms of a single outcome. After a statistical test is performed, investigators (and readers) can arrive at one of two conclusions:

- 1) The pattern is probably not due to chance (i.e., in common jargon, “There was a significant difference” or “The study was positive”).
- 2) The pattern is likely due to chance (i.e., in common jargon, “There was no significant difference” or “The study was negative”).

No matter how well the study is performed, either conclusion may be wrong. As shown in the Table below, a mistake about the first conclusion is labeled a type I error and a mistake about the second is labeled a type II error.

STUDY CONCLUSION	“TRUTH”	
	DIFFERENCE	NO DIFFERENCE
“Positive” study (significant difference)	True positive	Type I error
“Negative” study (no significant difference)	Type II error	True negative

Note that a type I error is only possible in a positive study, and a type II error is possible only in a negative study. Thus, this is one of the few areas of medicine where you can only make one mistake at a time.

### Type I Errors

A type I error is analogous to a false-positive result during diagnostic testing: A difference is shown when in “truth” there is none. Researchers have long been concerned about making this mistake and have conventionally demanded that the probability of a type I error be less than 5%. This convention is operationalized in the familiar critical threshold for  $P$  values:  $P$  must be less than 0.05 before we conclude that a study is positive. This means we are willing to accept that in 100 positive studies, at most 5 will be due to chance alone. The proba-

bility that a type I error has occurred in a positive study is the exact  $P$  value reported. For example, if the  $P$  value is 0.001, then the probability that the study has yielded false-positive results is 1 in 1000.\*

### Type II Errors

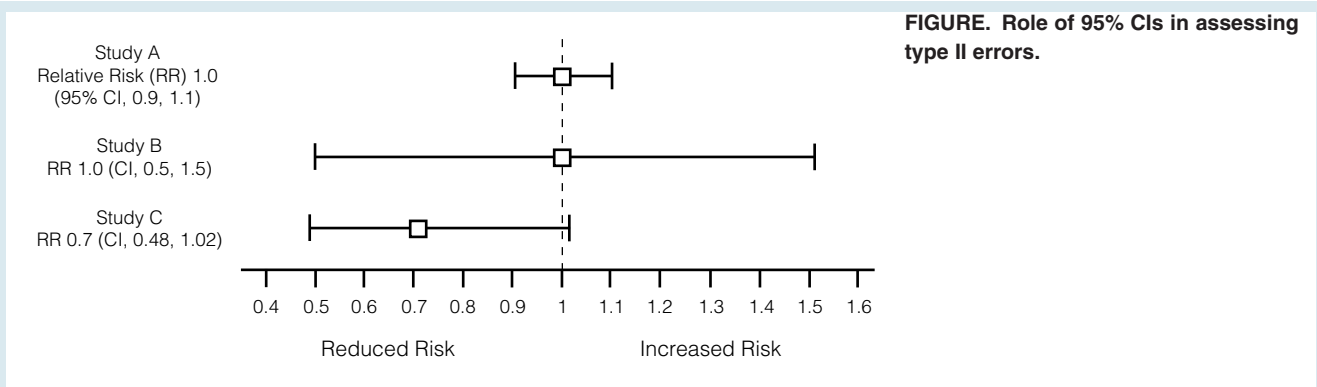
A type II error is analogous to a false-negative result during diagnostic testing: No difference is shown when in “truth” there is one. Traditionally, this error has received less attention from researchers than type I error and, consequently, may occur more often. Type II errors are generally the result of a researcher studying too few participants. To avoid the error, some researchers perform a sample size calculation before beginning a study and, as part of the calculation, assert what a “true difference” is and accept that they will miss it 10% to 20% of the time (i.e., type II error rate of 0.1 or 0.2). Regardless of how a study was planned, when faced with a negative study readers must be aware of the possibility of a type II error. Determining the likelihood of such an error is not a simple calculation but a judgment.

### Role of 95% CIs in Assessing Type II Errors

The best way to decide whether a type II error exists is to ask two questions: 1) Is the observed effect clinically important? and 2) To what extent does the confidence interval include clinically important effects? The more important the observed effect and the more the confidence interval includes important effects, the more likely that a type II error exists.

To gain some experience with this approach, consider the confidence intervals from three hypothetical randomized trials in the Figure. Each trial addresses the efficacy of an intervention to prevent a localized cancer from spreading. The outcome is the relative risk (RR) of metastasis (ratio of the risk in the intervention group over the risk in the control group). The interventions are not trivial, and you assert that you only consider risk reductions of greater than 10% to be clinically important. Note that each confidence interval includes 1—that is, each study is negative. There are no “significant differences” here. Which study is most likely to have a type II error?

\*This statement only considers the role of chance. Readers should be aware, however, that observed patterns may also be the result of bias.



**FIGURE.** Role of 95% CIs in assessing type II errors.

Study A suggests that the intervention has no effect (i.e. the relative risk is 1) and is very precise (i.e., the confidence interval is narrow). You can be confident that it is not missing an important difference. In other words, you can be confident that there's no type II error.

Study B suggests that the intervention has no effect (i.e., the RR is 1) but is very imprecise (i.e., the confidence interval is wide). This study may be missing an important difference. In other words, you should be worried about type II error, but this study is just as likely to be missing an important harmful effect as

an important beneficial one. A type II error is possible, and it could be in either direction.

Study C suggests that the intervention has a clinically important beneficial effect (i.e., the RR is much less than 1) and is also very imprecise. Most of the confidence interval includes clinically important beneficial effects. Consequently, a type II error is very likely. This is a study you would like to see repeated using a larger sample.

## Primer on Absolute vs. Relative Differences

When presenting data comparing two or more groups, researchers (and reporters) naturally focus on differences. Compared with others, one group may (pick one): cost more, have longer hospital stays, or have higher complication rates. These relations may be expressed as either absolute or relative differences. An absolute difference is a subtraction; a relative difference is a ratio. Because this choice may influence how big a difference “feels,” readers need to be alert to the distinction.

When the units are counts, such as dollars, the distinction between absolute and relative differences is obvious: group 1 costs \$30,000 more; group 1 had 40% higher costs. But when the units are percentages (frequently used to describe rates, probabilities, and proportions), it can be difficult to determine whether a stated difference is absolute or relative.

Consider the risk for blindness in a patient with diabetes over a 5-year period. If the risk for blindness is 2 in 100 (2%) in a group of patients treated conventionally and 1 in 100 (1%) in patients treated intensively, the absolute difference is derived by simply subtracting the two risks:

$$2\% - 1\% = 1\%$$

Expressed as an absolute difference, intensive therapy reduces the 5-year risk for blindness by 1%.

The relative difference is the ratio of the two risks. (NB: Relative risk, relative rate, rate ratios, and odds ratios are all examples of relative differences.) Given the data above, the relative difference is:

$$\frac{1\%}{2\%} = 50\%$$

Expressed as a relative difference, intensive therapy reduces the risk for blindness by half.

Both expressions have their place. Without any qualification, both statements (“reduced the risk by 1%” and “reduced the risk by 50%”) could be construed as representing either an absolute or relative difference. But most important, note the difference in “feel.” A statement of “reduced the risk by 1%” does feel like a smaller effect than “reduced the risk by 50%.”

The most frequent problem readers will face is the reporting of an isolated relative difference. Research abstracts, medical review articles, and general circulation newspapers and magazines are filled with statements like “60% decrease in costs,” “twice as many days in the hospital,” or “20% decrease in mortality.” These statements provide no information about the starting point. For example, the statement, “The risk for disease X was cut in half” gives no information about where you started. As shown in the Table below, there is a wide range of risks that can be cut in half.

Consequently, when you're

RISK FOR DISEASE		ABSOLUTE DIFFERENCE [A - B]	RELATIVE DIFFERENCE [B/A]
GROUP A	GROUP B		
20% (2/10)	10% (1/10)	10%	50%
2% (2/100)	1% (1/100)	1%	50%
0.2% (2/1000)	0.1% (1/1000)	0.1%	50%

presented with a relative difference (“60% more”) and you really want to get a complete picture of what’s going on, make sure you ask the question, “From what?” If the goal is clarity, the actual data (the dollars, the hospital days, and the mortality rates) for each group is tough to beat.

## Primer on Correlation Coefficients

Researchers are often interested in how two continuous variables relate to one another. To examine the relationship between body mass and fasting blood sugar, for example, one might study 20 people and measure both variables in each. The simplest approach to examine the relationship is to draw a picture, a scatterplot (an x-y graph), of body mass vs. fasting blood sugar. In this case, there are 20 dots, each representing one person.

Scatterplots of other relationships may involve different units of analysis, as shown in Table 1.

Any of these relationships can also be quantified by a single number—the correlation coefficient, also known as  $r$ . Because journals frequently only publish the number (and not the picture), this primer offers three questions to help readers visualize and interpret correlation coefficients.

TABLE 1

VARIABLE 1	VARIABLE 2	UNIT OF ANALYSIS
Body mass	Fasting blood sugar	Individual
Pneumococcal vaccination compliance	Years in practice	Physician practice
Mammography compliance	Pap smear compliance	Clinic
Physicians per capita	Death rate	State

### What Is the Sign on the Coefficient?

The first step is to look at the sign on  $r$ . If  $r$  is a positive number, the variables are directly related. In other words, as one goes up, so does the other (height and weight are a good example). If  $r$  is a neg-

ative number, the variables are inversely related. In other words, as one goes up, the other goes down (an example might be age and exercise capacity in adults). Knowing the sign helps you visualize the slope in the scatterplot, as shown in Figure 1.

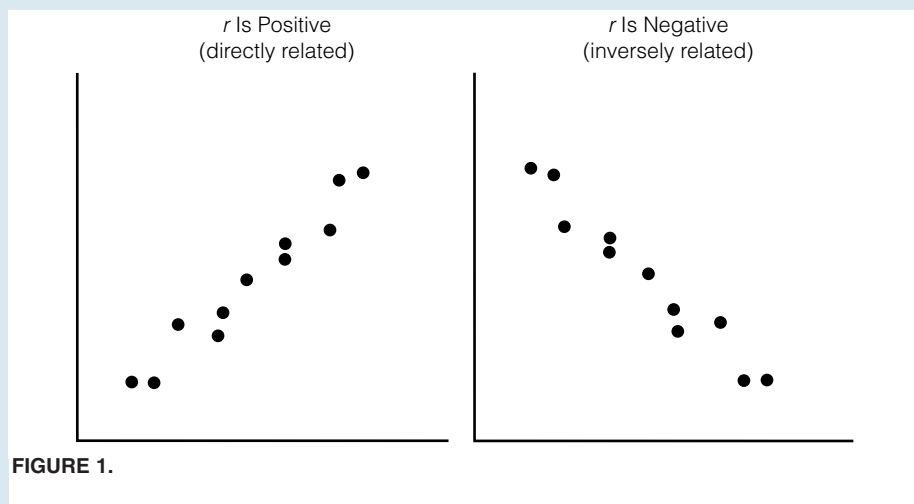


FIGURE 1.

### What Is the Magnitude of the Coefficient?

The next step is to consider how big  $r$  is;  $r$  ranges from  $-1$  to  $1$ . An  $r$  of  $0$  signifies absolutely no correlation, whereas an  $r$  of  $-1$  or  $1$  signifies a perfect correlation (all the data points fall on a line). In practice,  $r$  always has some intermediate value—there's always some correlation between two variables, but it's never perfect. The bigger the absolute value of  $r$  (i.e., the closer to  $-1$  or  $1$ ), the

stronger the correlation. The smaller the absolute value (i.e., the closer to  $0$ ), the weaker the correlation.

To provide perspective on what various  $r$ 's look like, Figure 2 shows three positive correlation coefficients and their associated scatterplots. (The scatterplots for the negative correlation coefficients would simply be mirror images.) Note that it may be difficult to see a relationship when  $r$  is less than  $0.3$  (or greater than  $-0.3$ ).

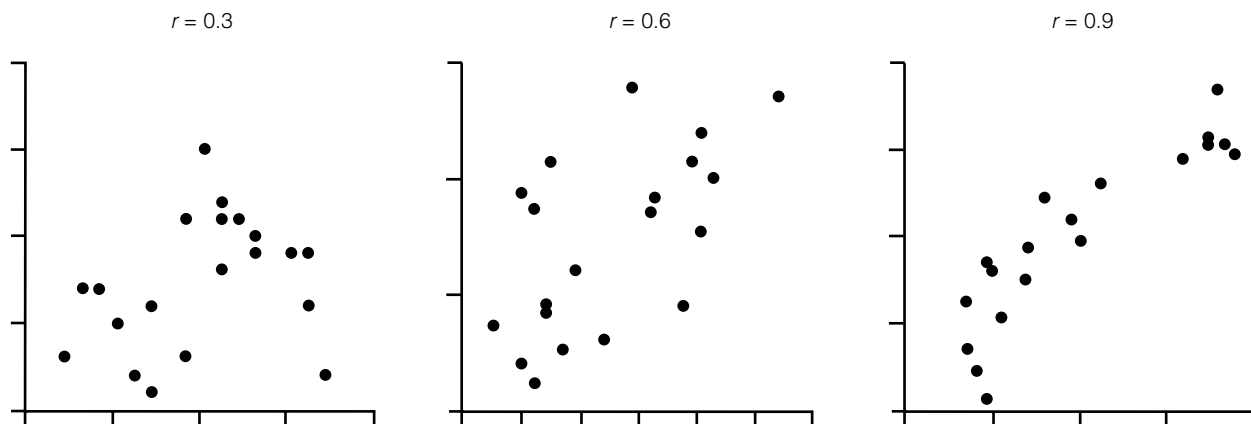


FIGURE 2.

The absolute magnitude of  $r$  is also a major determinant of statistical significance (the other being the number of observations). Consider 20 observations as depicted above. An  $r$  of 0.3 (a weak correlation) has an associated  $P$  value of 0.2. The  $P$  value falls with stronger correlations:  $P = 0.005$  for an  $r$  of 0.6 and  $P < 0.0001$  for an  $r$  of 0.9.

### Does the Coefficient Reflect a General Relationship or an Outlier?

A critical reader will want to consider if seeing a scatterplot might influence the interpretation of  $r$ . As shown in Figure 3, a single extreme data point (an outlier) can have a powerful effect on the correlation coefficient when the sample size is small.

To mitigate this problem,  $r$  is often recalculated substituting ranks for the raw data. (An  $r$  calculated using raw data is called a Pearson  $r$ , while an  $r$  calculated using ranks is called a Spearman  $r$ . A reported  $r$  should be assumed to be Pearson  $r$  unless otherwise noted.)

For example, fasting blood sugar levels of 610, 320, 290, and 280 mg/dL would be converted to ranks 1, 2, 3, and 4; body weights of 350, 270, 220, and 210 lb would be converted to ranks 1, 2, 3, and 4; and the data point (610, 220) would become (1, 3). This recalculation does not eliminate the effect of outliers, but it does help to dampen their effects (in Figure 3, from left to right the recalculated  $r$ 's are 0.56, 0.62, and 0.37). In small samples, this recalculation can be particularly important.

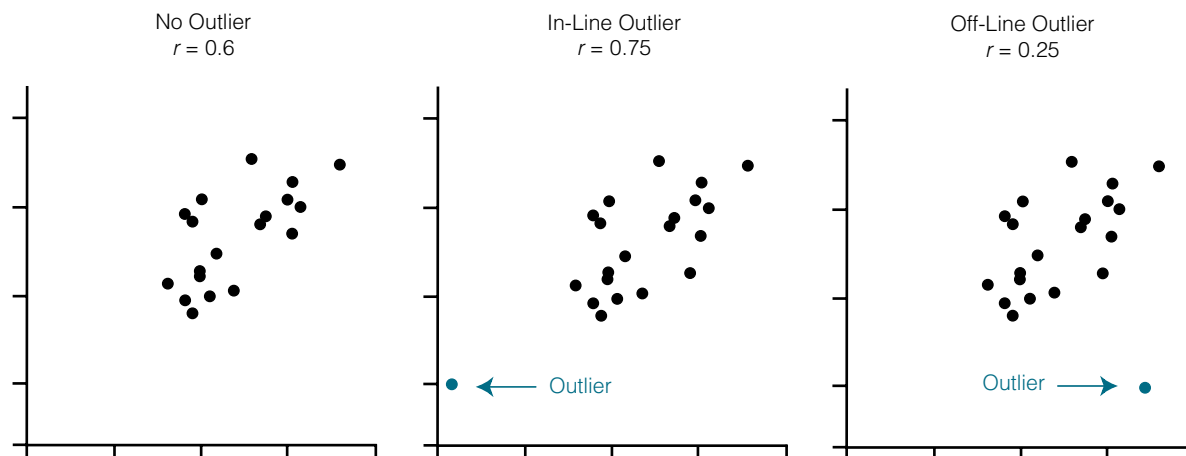


FIGURE 3.

Although correlation coefficients are an efficient way to communicate the relationship between two variables, they are not sufficient to interpret a relationship. The unit of analysis also matters. For example, a strong positive correlation between influenza and pneumococcal vaccination rates measured among physicians should be interpreted differently than the same coefficients measured among clinics. The former may imply that physicians have different beliefs about vaccinations, whereas the latter may simply reflect that clinics differ in the resources devoted

to vaccination (e.g., reminder systems, nurse-run vaccination clinics).

Finally, correlation coefficients do not communicate information about whether one variable moves in response to another. There is no attempt to distinguish between the two variables—that is, to establish one as dependent and the other as independent. Thus, relationships identified using correlation coefficients should be interpreted for what they are: associations, not causal relationships.

*A compendium of ecp primers from past issues can be viewed and/or requested at <http://www.acponline.org/journals/ecp/primers.htm>.*

## Primer on 95% CIs for the Number Needed To Treat

Few, if any, therapeutic interventions benefit every patient. One way to gauge the likelihood that one patient will benefit is to calculate the number needed to treat (NNT) — that is, the number of patients who must be treated for one to benefit. The general approach is as follows:

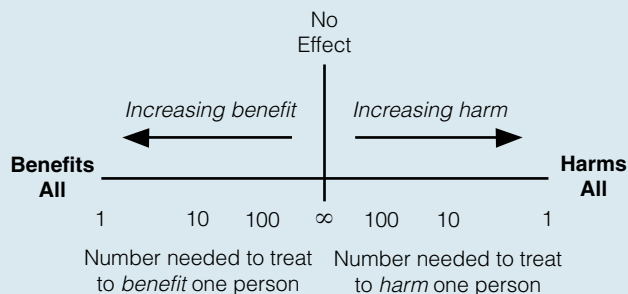
$$\frac{\text{Percentage with outcome}_{\text{standard treatment}} - \text{Percentage with outcome}_{\text{new treatment}}}{\text{Absolute risk reduction}}$$

$$100/\text{Absolute risk reduction} = \text{Number needed to treat}$$

For example, consider a randomized trial in which 50% of the participants die in the control group and 40% die in the intervention group. The absolute risk reduction for death is thus 10%, and the NNT to avoid a death is 10 (100/10)\*. This treatment would be preferred over a competing treatment whose NNT to avoid death was 20.

NNT can be calculated using any dichotomous outcome (an outcome that a patient either experiences or does not experience). In most cases, the NNT is calculated by using an adverse outcome — one that most persons would prefer to avoid (e.g., angina, myocardial infarction, cardiac death, any death). But because different outcomes are possible, an NNT of 10 is not always preferable to an NNT of 20 (e.g., if the former were for angina and the latter for any death). Therefore, an NNT should always be accompanied by a clearly specified outcome.

As is the case with all variables measured in research, the NNT is an estimate. The precision of the estimate is largely a function of how many people were studied and is reflected by using a 95% CI. The 95% CI for an NNT is the range of values in which we would expect to find the “true” NNT 95% of the time.† In some cases, the range may also include the possibility of harm. A 95% CI for an NNT that contains the possibility for both harm and benefit passes through infinity. In other words, an intervention with no effect has an NNT of infinity. This notion is probably most easily understood by considering the continuum of possible NNTs:



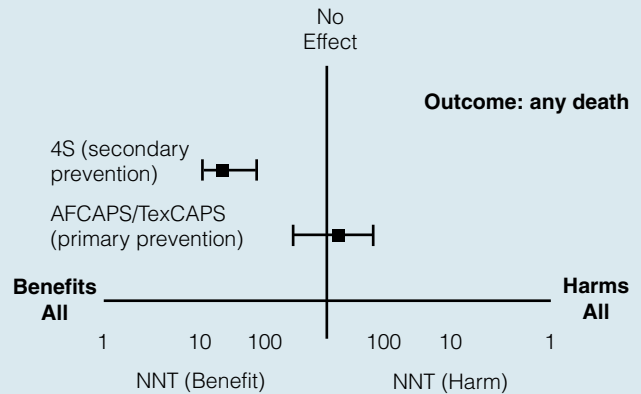
\*For readers who prefer decimals,  $NNT = 1/\text{Absolute risk reduction}$ . In this example, 1/0.1 or 10.

†Apologies to statistical purists who would direct the reader toward a more formal definition for a 95% CI: “The interval computed from the sample data which, were the study repeated multiple times, would contain the unknown parameter 95% of the time.”

95% CIs for NNTs that contain the possibility of both harm and benefit are probably best communicated graphically. Altman introduced the concept in a recent article in the *BMJ*,<sup>1</sup> and proposed the following labels: NNT (benefit) and NNT (harm).

The importance of a graphic display is best demonstrated by example. Consider 95% CIs for NNTs for lipid-lowering therapy. The outcome is death from any cause. In the Scandinavian Simvastatin Survival Study (4S)<sup>2</sup> (which studied simvastatin in patients who had either angina or previous myocardial infarction), the 95% CIs for NNTs did not pass through infinity. The NNT (benefit) was 30 (95% CI, 19 to 68). In the Air Force Coronary/Texas Atherosclerosis Prevention Study (AFCAPS/TexCAPS)<sup>3</sup> (which studied lovastatin in patients without heart disease who had normal cholesterol levels), however, the CI does pass through infinity. The NNT (harm) was 1130; 95% CI, NNT (benefit) 153 to ∞ to NNT (harm) 120. For most of us, these data would be better summarized in a figure:

NNT and the 95% CIs for NNT are relatively new concepts.



Whether they represent a genuine advance in communicating data to clinicians is unknown. As always, we are interested in your thoughts.

### References

1. Altman DG. Confidence intervals for the number needed to treat. *BMJ*. 1998;317:1309-12.
2. Scandinavian Simvastatin Survival Study Group. Randomized trial of cholesterol lowering in 4444 patients with coronary heart disease: the Scandinavian Simvastatin Survival Study (4S). *Lancet*. 1994;344:1383-9.
3. Down JR, Clearfield M, Weis S, et al. Primary prevention of acute coronary events with lovastatin in men and women with average cholesterol levels: results of AFCAPS/TexCAPS. *JAMA*. 1998;279:1615-22.

# Primer on Statistical Significance and P Values

In the world of medical journals, few phrases evoke more authority than “the differences observed were statistically significant.” Unfortunately, readers frequently accord too much importance to this statement and are often distracted from more pressing issues. This Primer reviews the meaning of the term *statistical significance* and includes some important caveats for critical readers to consider whenever it is used.

## Assessing the Role of Chance

Consider a study of a new weight loss program: Group A receives the intervention and loses an average of 10 pounds, while group B serves as a control and loses an average of 3 pounds. The main effect of the weight loss program is therefore estimated to be a 7-pound weight loss (on average). But we would rarely expect that any two groups would have exactly the same amount of weight change. So could it just be chance that group A lost more weight?

There are two basic statistical methods used to assess the role of chance: confidence intervals (the subject of next issue’s Primer) and hypothesis testing. As shown in the Figure below, both use the same fundamental inputs.

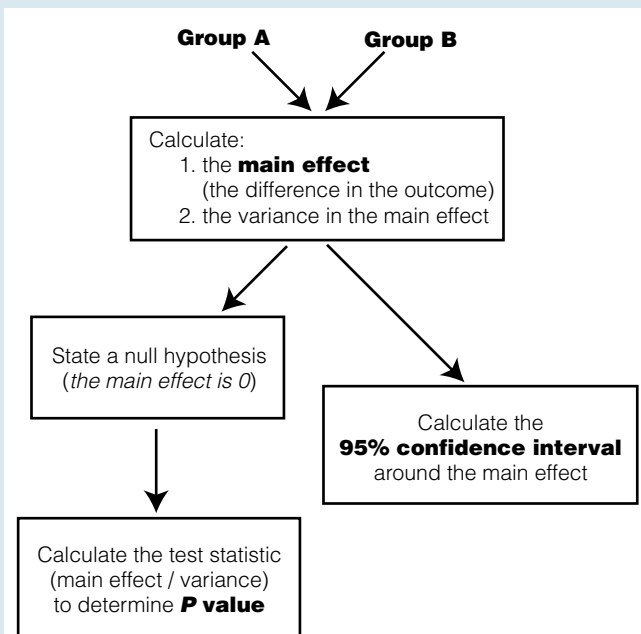


FIGURE 1. Statistical approach to comparing two groups.

Hypothesis testing goes on to consider a condition—the null hypothesis—that no difference exists. In this case, the null hypothesis is that the weight change in the two groups is the same. The test addresses the question, “If the true state of affairs

is no difference (i.e., the null hypothesis is true), what is the probability of observing this difference (i.e., 7 lbs) or one more extreme (i.e., 8 lbs, 9 lbs, etc.)”? This probability is called the *P* value and, for most of us, translates roughly to “the probability that the observed result is due to chance.”

If the *P* value is less than 5%, researchers typically assert that the findings are “statistically significant.” In the case of the weight loss program, if the chance of observing a difference of 7 pounds or more (when, in fact, none exists) is less than 5 in 100, then the weight loss program is presumed to have a real effect.

TABLE 1  
Relationship between Common Language and Hypothesis Testing

COMMON LANGUAGE	STATISTICAL STATEMENT	CONVENTIONAL TEST THRESHOLD
“Statistically significant” “Unlikely due to chance”	The null hypothesis was rejected.	$P < 0.05$
“Not significant” “Due to chance”	The null hypothesis could not be rejected.	$P > 0.05$

Table 1 shows how our common language relates to the statistical language of hypothesis testing.

## Factors That Influence P Values

Statistical significance (meaning a low *P* value) depends on three factors: the main effect itself and the two factors that make up the variance. Here is how each relates to the *P* value:

- *The magnitude of the main effect.* A 7-lb difference will have a lower *P* value (i.e., more likely to be statistically significant) than a 1-lb difference.
- *The number of observations.* A 7-lb difference observed in a study with 500 patients in each group will have a lower *P* value than a 7-lb difference observed in a study with 25 patients in each group.
- *The spread in the data (commonly measured as a standard deviation).* If everybody in group A loses about 10 pounds and everybody in group B loses about 3 pounds, the *P* value will be lower than if there is a wide variation in individual weight changes (even if the group averages remain at 10 and 3 pounds). Note: More observations do not reduce spread in data.

## Caveats about the Importance of P Values

Unfortunately, *P* values and statistical significance are often accorded too much weight. Critical readers should bear three facts in mind:

### 1. The $P < 0.05$ threshold is wholly arbitrary.

There is nothing magical about a 5% chance—it's simply a convenient convention and could just as easily be 10% or 1%. The arbitrariness of the 0.05 threshold is most obvious when *P* values are near the cut-off. To call one finding significant when the *P* value is 0.04 and another not significant when it is 0.06 vastly overstates the difference between the two findings.

Critical readers should also realize that dichotomizing *P* values into simply "significant" and "insignificant" loses information in the same way that dichotomizing any clinical laboratory value into "normal" and "abnormal" does. Although serum sodium levels of 115 and 132 are both below normal, the former is of much greater concern than the latter. Similarly, although both are significant, a *P* value of 0.001 is much more "significant" than a *P* value of 0.04.

### 2. Statistical significance does not translate into clinical importance.

Although it is tempting to equate statistical significance with clinical importance, critical readers should avoid this temptation. To be clinically important requires a substantial change in an outcome that matters. Statistically significant changes, however, can be observed with trivial outcomes. And because significance is powerfully influenced by the number of observations, statistically significant changes can be observed with trivial changes in important outcomes. As shown in Table 2, large studies can be significant without being clinically important and small studies may be important without being significant.

### 3. Chance is rarely the most pressing issue.

Finally, because *P* values are quantifiable and seemingly objective, it's easy to overemphasize the importance of statistical significance. For most studies, the biggest threat to an author's conclusion is not random error (chance), but systematic error (bias). Thus, readers must focus on the more difficult, qualitative questions: Are these the right patients? Are these the right outcomes? Are there measurement biases? Are observed associations confounded by other factors?

**TABLE 2**  
**Big Studies Make Small Differences "Significant"\***

SIZE, <i>n</i> (IN EACH GROUP)	WEIGHT LOSS		MAIN EFFECT	P VALUE	APPROPRIATE CONCLUSION
	GROUP A (INTERVENTION)	GROUP B (CONTROL)			
10	20 lb	3 lb	17 lb	0.07	Not significant, but promising
1000	5 lb	3 lb	2 lb	0.03	Significant, but clinically unimportant

\*The standard deviation of the weight change is assumed to be 20 lb.

A compendium of **ecp** primers from past issues can be viewed and/or requested at <http://www.acponline.org/journals/ecp/primers.htm>.

# Primer on 95% Confidence Intervals

Readers frequently face questions about the role of chance in a study's results. The traditional approach has been to consider the probability that an observed result is due to chance—the *P* value. However, *P* values provide no information on the results' precision—that is, the degree to which they would vary if measured multiple times. Consequently, journals are increasingly emphasizing a second approach: reporting a range of plausible results, better known as the 95% confidence interval (CI). This Primer reviews the concept of CIs and their relationship to *P* values.

## Assessing the Role of Chance

There are two basic statistical methods used to assess the role of chance: hypothesis testing (which results in a *P* value—the subject of last issue's Primer) and 95% CIs. As shown in Figure 1, both use the same fundamental inputs.

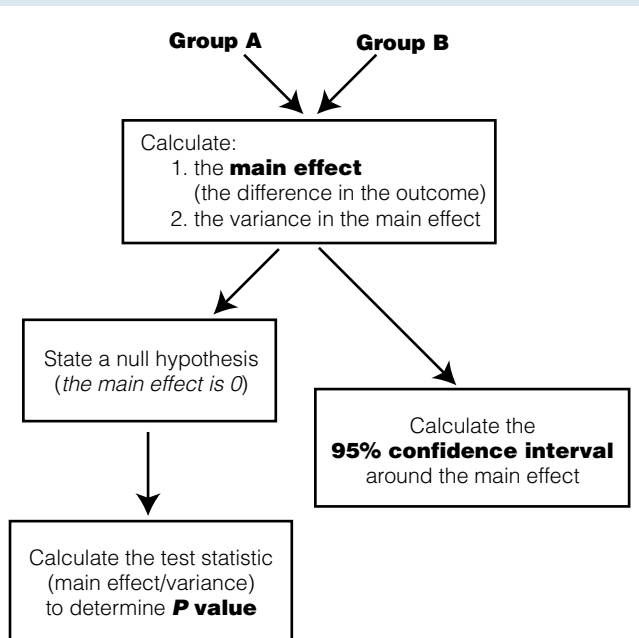


FIGURE 1. Statistical approach for comparing two groups.

Consider a study of a new weight loss program: Group A receives the intervention and loses an average of 10 pounds, whereas group B serves as a control and loses an average of 3 pounds. The main effect of the weight loss program is therefore estimated to be a 7-pound weight loss (on average).

But readers should recognize that the true effect of the program may not be exactly a 7-pound weight loss. Instead, the true effect is best represented as a range. What is the range of effects that might be expected just by chance? That is the ques-

tion addressed by a 95% CI. In this example the study abstract might read:

The mean weight loss was 10 pounds for patients in the intervention group and 3 pounds for patients in the control group, resulting in a mean difference of 7 pounds and a 95% CI of 2 to 12. In other words, 95% of the time the true effect of the intervention will be within the range from 2 to 12 pounds.

To conceptualize the more formal definition of a 95% CI, it is useful to consider what would happen if the study were repeated 100 times. Obviously, not every study would result in a 7-pound weight loss in favor of the intervention. Simply due to the

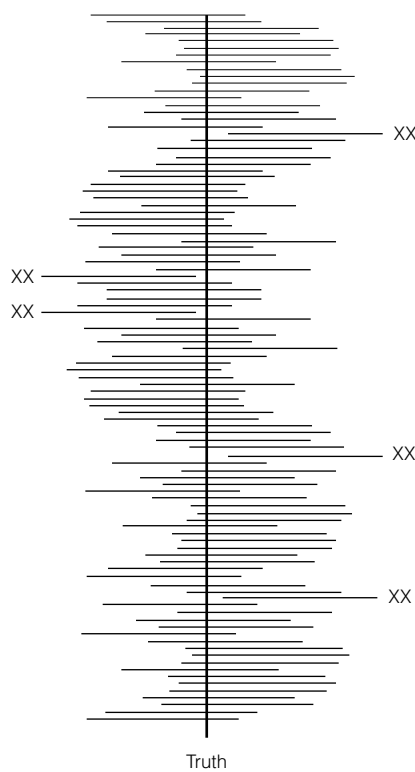
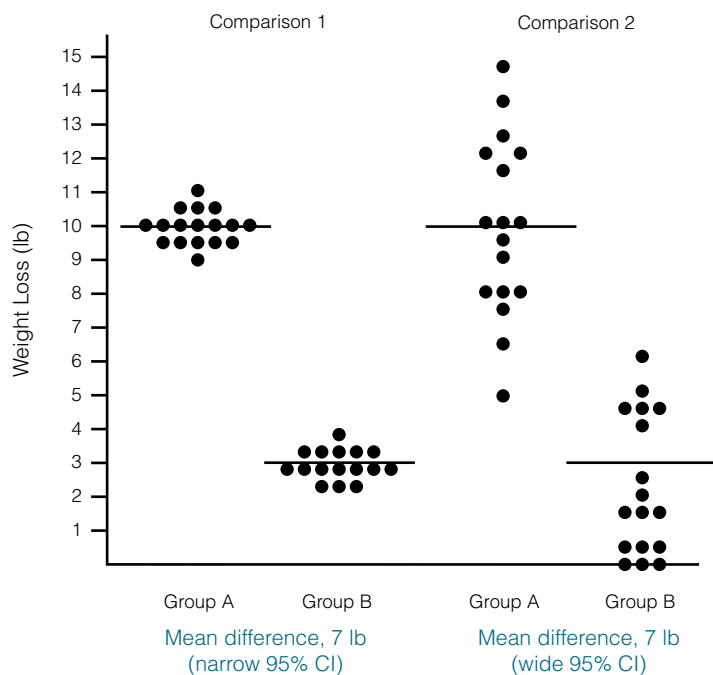


FIGURE 2. Every study can have a 95% CI.

play of chance, weight loss would be greater in some studies and less in others, and some studies might show that the controls lost more weight. As shown in Figure 2, we can generate a 95% CI for each study.

Note that for 95 out of 100 studies, the CI contains the truth (and 5 times out of 100 it does not). This example helps explain the formal definition of a 95% CI: "The interval computed from the



**FIGURE 3.** More diversity in weight loss equals a larger 95% CI. The horizontal bars represent group means.

sample data which, were the study repeated multiple times, would contain the true effect 95% of the time.”

### Factors That Influence 95% CIs

Confidence intervals really are a measure of how precise an estimated effect is. The range of a CI is dependent on the two factors that cause the main effect to vary:

1) *The number of observations.* This factor is largely under the investigator’s control. A 7-pound difference observed in a study with 500 patients in each group will have a narrower CI than a 7-pound difference observed in a study with 25 patients in each group.

2) *The spread in the data* (commonly measured as a standard deviation). This factor is largely outside the investigator’s control. Consider the two comparisons in Figure 3. In both cases, the mean weight loss in group A is 10 pounds and the mean weight loss in group B is 3 pounds. If everybody in group A loses

about 10 pounds and everybody in group B loses about 3 pounds, then the CI will be narrower (left part of figure) than if individual weight changes are spread all over the map (right part of figure).

Readers will occasionally encounter CIs calculated for other confidence levels (e.g., 90% or 99%). The higher the degree of confidence, the wider the confidence interval. Thus, a 99% CI for the 7-pound difference would have to be wider than the 95% CI for the same data.

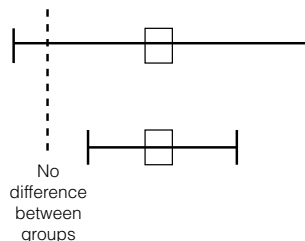
### Relationship between 95% CIs and P values

Information about the *P* value is contained in the 95% CI. As shown in Figure 4, the *P* value can be inferred based on whether the finding of “no difference” falls within the CI.

So, given a CI of 2 to 12 pounds for the 7-pound difference, one could infer that the *P* value is less than 0.05. Alternatively,

If the 95% CI includes no difference between groups, then the *P* value is  $> 0.05$ .

If the 95% CI does not include no difference between groups, then the *P* value is  $< 0.05$ .



**FIGURE 4.** Relationship between *P* value and 95% CI.

given a CI of -3 to 17 pounds for the 7-pound difference, one could infer that the *P* value is greater than 0.05. If the CI terminates exactly on no difference, such as 0 to 14 pounds, then the *P* value is exactly 0.05.

Remember that the value for no difference depends on the type of effect measure used. When the effect measure involves a subtraction, the value for the difference is 0. When the effect measure involves a ratio, the value for no difference is 1. As shown in Table 1, readers must pay careful attention to this in order to reliably interpret the CI.

Although *P* values and 95% CIs are related, CIs are preferred because they convey information about the range of plausible effects. In other words, the CI provides the reader with some sense of how precise the estimate of the effect is. This is a valuable dimension that is not contained within a *P* value.

But, like *P* values, 95% CIs do not answer two critical questions: 1) Is the result correct? 2) Is the observed effect “important”? To answer the first question, readers must seek other data and evaluate the possibility of systematic error (bias). To answer the second, they must rely on their own clinical judgment.

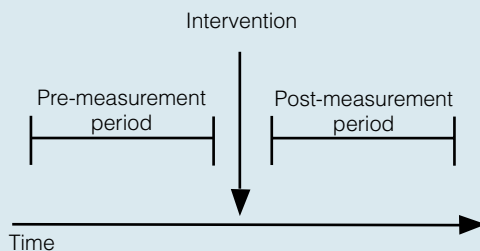
**TABLE 1**  
**Examples Demonstrating 95% CIs and P Values**

EXAMPLE	EFFECT MEASURE	VALUE FOR NO DIFFERENCE	CI INCLUDES NO DIFFERENCE?	STATISTICALLY SIGNIFICANT? ( <i>P</i> < 0.05)
The average weight loss was 7 lbs (95% CI, -3 to 17)	Difference in means	0	Yes	No
42% absolute reduction in the need for intubation (95% CI, 7% to 70%)	Difference in proportions	0	No	Yes
The relative risk for cancer was 2.3 for smokers compared with nonsmokers (95% CI, 1.8 to 3.0)	Relative risk	1	No	Yes
The odds ratio for readmission was 0.8 for managed care patients (95% CI, 0.3 to 1.2)	Odds ratio	1	Yes	No

*A compendium of ecp primers from past issues can be viewed and/or requested at <http://www.acponline.org/journals/ecp/primers.htm>.*

## A Primer on Before-After Studies: Evaluating a Report of a “Successful” Intervention

It can be difficult to rigorously evaluate a clinical management or quality improvement intervention. Because these interventions generally occur at a system level (i.e., throughout the clinic, the hospital, or the health plan), it may not be practical to obtain suitable concurrent controls (clinics, hospitals, or plans not exposed to the intervention). As illustrated below, a common approach is to measure outcomes before the intervention is implemented and compare them with outcomes measured afterward—an approach often called a *before-after study* (or a *pre-post study*).



Although academics can easily criticize the lack of a concurrent control group, managers still need to make decisions on the basis of data available to them. This primer is intended to provide guidance on how to think critically about a report of a “successful” intervention obtained from a before-after study.

As with any report of “success,” readers should start by asking three questions: Is the outcome unimportant? Is the magnitude of the change trivial? Were critical outcomes ignored? If the reader is comfortable that the answer to each is no, then he or she must go on to challenge the fundamental inference: that the “success” is a consequence of the intervention. The validity of this inference is threatened with an affirmative response to any of the following questions:

**Would all participants in the “before group” be eligible for the “after group”?** A typical before-after study compares the outcomes of hospitalized patients before and after some system intervention. Thus, different patients are often involved (e.g., patients admitted with pneumonia in June are compared with patients admitted with pneumonia in July). If only certain patients are eligible for the intervention, however, an inference about the success of the intervention can be seriously flawed. Consider a study of the effect an outpatient low-molecular-weight heparin program (which, by necessity, excludes the sickest patients) on the average length of stay of patients with

deep venous thrombosis (DVT). A comparison of cost between all patients who have DVT (before) and patients who have DVT and are eligible for the outpatient program (after) would dramatically overestimate the effect of the intervention. The best estimate of the intervention’s effect would be to compare all patients with DVT (before) with all patients with DVT (after), including both those who are eligible and those who are ineligible for the program. The comparability of patients in the before group and the after group is particularly relevant in assessments of the effect of guidelines (which generally apply to select patient subgroups).

**Is there evidence for a prevailing “temporal trend”?** Many outcomes change over time, regardless of whether an intervention has been applied. Consider a before-after study testing an intervention to reduce length of stay in the hospital. The average length of stay is 5 days before the introduction of the intervention but is 4.7 days after introduction. It is tempting to believe that the intervention caused the change. On the other hand, there is a prevailing temporal trend: Length of stay has been decreasing everywhere across time (at least until recently). The same problem would arise in a before-after study that tested an intervention to increase the use of aspirin in patients who have had a myocardial infarction. It would be difficult to untangle whether the observed change is the result of the intervention or dramatic television advertising. Because many forces are likely to be acting on outcomes that people care about, it is important to question whether an intervention is truly responsible for “success,” particularly if outcomes are improving everywhere.

**Were study participants selected because they were “outliers”?** Understandably, some before-after studies target “problem areas” and select persons who are “outliers”—that is, participants who have extreme values in some measure. These studies may follow the same participants over time and face another threat to validity: regression to the mean. Examples could include a study of case management in patients who have had high utilization in the past or a study of an intensive communication tutorial in physicians who have been judged by their patients to have poor communication skills. Even if there is no intervention, participants selected because of extreme values will, on average, be found to have less extreme values with repeated measurement. Extremely high utilization in 1 year tends not to be so high the next (some patients may have had a major heart attack, stroke, or other catastrophic event that does not occur again in the next year); a group of physicians with extremely poor communication skills will tend to improve (some may have had a personal crisis that resolves in the ensuing year). Note that in neither case are the participants expected to return to the mean; they just become less extreme. Regression to the mean sets the stage to ascribe changes to a case management program or a communication tutorial when they actually represent the natural course of events.

Although it is always possible that a change observed in a before–after study is a consequence of the intervention, affirmative responses to any of the preceding questions make the inference more tenuous. Alternatively, the inference is strengthened when investigators paid careful attention to the comparability of the participants. Inferences are further strengthened when the observed change is substantial, unique, and occurs quickly after the intervention—in other words, when it is diffi-

cult to ascribe the finding to temporal trends. The confusing effect of regression to the mean can be avoided if participants are not selected because they are outliers. Nonetheless, inferences from a before–after study should be seen as being based on circumstantial evidence. If the accuracy of the inference is important, readers and researchers alike must ask whether there is a reasonable opportunity to test the intervention by using concurrent controls.

# Primer on Group Randomized Trials

Group randomized trials are experiments in which the intervention occurs at the level of the group (typically physicians or clinics) but observations are made on individuals within the groups (e.g., patients). Because group randomized trials are increasingly common in health services research, critical readers should understand their rationale, the implications of group size vs. number of groups, and the limitations of the approach.

## Why Randomize by Group?

Group randomization is particularly useful when there is a high risk for contamination if group members are randomized as individuals. For example, an investigator studying the effects of a clinical practice guideline can't assume that a provider caring for patients in the intervention arm will not apply this knowledge to the patients assigned to the control arm. Such contamination biases the study toward a finding of no effect. Randomizing at the level of the physician avoids this source of contamination because physicians are either exposed or not exposed to the intervention. If there are concerns that intervention physicians will contaminate control physicians in the same clinic, randomization should occur at the clinic level.

## Group Size vs. Number of Groups

To illustrate some of the issues raised by group randomization, consider a trial to test a cholesterol management guideline. Physicians would be randomly assigned to a control or an intervention arm while the outcome (say, the mean change in cholesterol after 6 months) would be measured on their patients. As shown in the Figure, however, there are many possible combinations of group size and number of groups.

In each case we have 200 patient observations (100 patients in each arm), but as group size increases there are fewer physicians. With smaller group size, there is less information on many physicians; with larger group size, there is more information on only a few physicians. Because the study is intended to measure the impact of the guideline on physicians, the design

with 40 physicians is more likely to detect a significant intervention effect than the one with only 8 physicians—despite the equivalent size of the patient sample. In other words, collecting a large amount of information on patients in one physician practice allows something precise to be said about that physician but adds little to the ability to answer the study question.

Although ideally there should be as many physicians as possible, practical considerations often limit enrollment. The number of physicians available and willing to participate is often limited. It can be very expensive to enroll and train a physician. It is often easier to recruit many patients and a few physicians than it is to recruit many physicians. Thus, there is a trade-off between increasing group size (often the most expedient way to increase sample size) and increasing the number of groups (generally the most effective way to increase power).

## Sample Size in Group Randomized Trials

The ability to make statistical inferences is inversely related to variability in the outcome measure. In this example, the variability in cholesterol can come from two sources: differences among patients and differences among physicians (presumably in their ability to influence the patients' cholesterol either through behavior modification or pharmacologic treatment). The *proportion* of cholesterol variability attributable to physicians is called the intraclass correlation (the term is a misnomer because it has nothing to do with correlation), or rho. As rho increases, a greater share of the variability comes from physicians, so that increasing the number of physicians will become more important. If rho is small, then increasing the number of patients per physician may be sufficient to increase the power to detect an effect. Rho can only be zero if there is no systematic difference between groups. In other words, 1) physicians do not differ in their response to education and 2) the patients of one physician do not differ systematically from those of another. A typical rho in this setting is between 0.01 and 0.04.

Table 1 illustrates how changes in the intraclass correlation affect the sample size needed to produce equivalent levels of

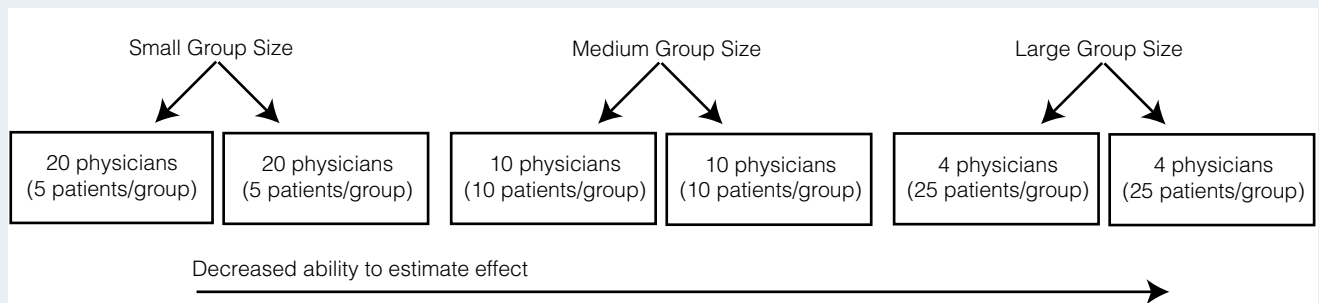


FIGURE. The relationship between group size and the number of groups. Size for each group is 200.

TABLE 1

**Relationship between Intraclass Correlation, Sample Size, and Number of Groups**

NUMBER OF PHYSICIANS PER TREATMENT ARM	INTRACLASS CORRELATION (RHO)	SAMPLE SIZE (TOTAL)
4	0.0*	200
	0.01	396
	0.02	$\infty$
	0.03	$\infty$
10	0.0*	200
	0.01	248
	0.02	320
	0.03	486
20	0.0*	200
	0.01	220
	0.02	246
	0.03	278

\*No physician effect.

precision. As the intraclass correlation increases, the total number of patients needed also increases. In addition, Table 1 shows how the effect is modified by the number of physicians. When the intraclass correlation is 0.03, for example, a study with 10 physicians in each arm requires 486 patients to achieve the same precision as a study with 278 patients and 20 physicians in each arm. Notice that with 4 physicians in each arm, no number of patients would provide sufficient information to answer the study question. This illustrates a major limitation of group randomized trials: It may be impossible to collect enough data at the patient level to make up for a small number of groups. The important lesson here is that the *effective* sample size in a group randomized trial is not related only to the number of patients but depends on the number of groups and the intraclass correlation.

**Comparability of Patients**

One of the most important advantages of randomization is that, if the trial is large enough, it is fair to assume that the study groups will be comparable with respect to all variables (measured and unmeasured). This enhances our ability to make inferences about the effect of the intervention on the outcome. In contrast to randomized trials of individuals, group randomized trials involve only a limited number of groups—typically 15 or 20. Thus, there are rarely enough groups to ensure even distribution of variables that could confound the treatment effect and bias the outcomes comparison.

As a result, investigators need to collect information on important confounders and plan analyses that will control for these factors. These analyses require special techniques that directly incorporate the group structure (cluster analyses). It would be a mistake in our hypothetical example to simply compare the average cholesterol levels in the treatment and control group with, say, a standard z-test. For example, a study with  $\rho = 0.03$ , 10 physicians per group, and 486 total patients would be equivalent to a study with  $\rho = 0$  and 200 total patients. A z-test would calculate a standard error based on 486 patients, when the effective sample size is only 200. Statistical analysis that ignores this fact can give falsely low *P* values and overly optimistic confidence intervals.

Policymakers and managers are increasingly interested in moving “hard science” to the vagaries of actual clinical practice. To help translate efficacy into effectiveness, interventions are being directed to physicians (or groups of physicians). Group randomization is the best approach to make valid inferences about their value.

*This Primer was contributed by Michael L. Beach, MD, PhD, Dartmouth–Hitchcock Medical Center, Lebanon, New Hampshire.*

*A compendium of ecp primers from past issues can be viewed and/or requested at <http://www.acponline.org/journals/ecp/primers.htm>.*

# Primer on Cost-Effectiveness Analysis

Cost-effectiveness analysis (CEA) is a technique for selecting among competing wants wherever resources are limited. Developed in the military, CEA was first applied to health care in the mid-1960s and was introduced with enthusiasm to clinicians by Weinstein and Stason in 1977:

“If these approaches were to become widely understood and accepted by the key decision makers in the health-care sector, including the physician, important health benefits or cost savings might be realized.”

Regardless of whether this hope was realized, CEA has since become a common feature in medical literature.

## The Basics of CEA

CEA is a technique for comparing the relative value of various clinical strategies. In its most common form, a new strategy is compared with current practice (the “low-cost alternative”) in the calculation of the cost-effectiveness ratio:

$$\text{CE ratio} = \frac{\text{cost}_{\text{new strategy}} - \text{cost}_{\text{current practice}}}{\text{effect}_{\text{new strategy}} - \text{effect}_{\text{current practice}}}$$

The result might be considered as the “price” of the additional outcome purchased by switching from current practice to the new strategy (e.g., \$10,000 per life year). If the price is low enough, the new strategy is considered “cost-effective.”

It’s important to carefully consider exactly what that statement means. If a strategy is dubbed “cost-effective” and the term is used as its creators intended, it means that the new strategy is a good value. Note that being cost-effective does not mean that the strategy saves money, and just because a strategy saves money doesn’t mean that it is cost-effective. Also note that the very notion of cost-effective requires a value judgment—what you think is a good price for an additional outcome, someone else may not.

It’s also worthwhile to recognize that CEA is only relevant to certain decisions. Table 1 delineates the various way a new

**TABLE 1**  
**Conditions under Which CEA Is Relevant**

EFFECTIVENESS	COST	
	NEW STRATEGY COSTS MORE	NEW STRATEGY COSTS LESS
New strategy is <i>more</i> effective	CEA relevant	Adopt new strategy
New strategy is <i>less</i> effective	New strategy is “dominated”	CEA relevant

strategy might compare with an existing approach. Note that a CEA is relevant only if a new strategy is *both* more effective and more costly (or both less effective and less costly).

## An Example

Consider two strategies intended to lengthen life in patients with heart disease. One is simple and cheap (e.g., aspirin and  $\beta$ -blockers); the other is more complex, more expensive, and more effective (e.g., medication plus cardiac catheterization, angioplasty, stents, and bypass). For simplicity, we will assume that doing nothing has no cost and no effectiveness. Table 2 shows the relevant data.

Note that CEA is about marginal (also called incremental) costs and benefits. So the marginal cost of a simple strategy is the difference between the cost of that strategy and the cost of doing nothing. The marginal cost for the complex strategy is the difference between the cost of the complex strategy and the cost of the simple strategy (not the cost of doing nothing). The calculation is similar for effectiveness. The final outcome measure for the analysis is the CE ratio: the ratio of marginal cost to marginal effectiveness.

**TABLE 2**  
**A CEA Examining Three Strategies**

STRATEGY	COST	MARGINAL COST	EFFECTIVENESS	MARGINAL EFFECTIVENESS	CE RATIO
Nothing	\$0	—	0 years	—	—
Simple	\$5000	\$5000	5 years	5 years	\$1000/yr
Complex	\$50,000	\$45,000	5.5 years	0.5 years	\$90,000/yr

TABLE 3

## A CEA Examining Two Strategies

STRATEGY	COST	MARGINAL COST	EFFECTIVENESS	MARGINAL EFFECTIVENESS	CE RATIO
Nothing	\$0	—	0 years	—	—
Complex	\$50,000	\$50,000	5.5 years	5.5 years	\$9091/yr

### Things To Ask

If a study is of interest and its primary outcome is a cost-effectiveness ratio, critical readers should seek answers to the following questions.

#### 1. Are the relevant strategies being compared?

Because CEA involves marginal cost and benefits, the choice of which strategies to compare can drive the calculation and the conclusion of a CEA. Consider the effect of repeating the above analysis *without* the simple strategy (Table 3).

By excluding the simple strategy, the CE ratio for the complex strategy falls from \$90,000 per life-year to \$9091 per life-year.

Thus, CEA is very sensitive to the choice of strategies being compared. Readers need to carefully consider whether the choice being presented is really the choice that interests clinicians.

#### 2. How good are the effectiveness data?

It's hard to get too excited about cost-effectiveness if the effectiveness of the strategy is really unknown. So as a first step, the critical reader should examine the information used for effectiveness. Ideally, the data should come from randomized trials. If they don't, you'll want to scrutinize the face validity of the assumptions. Unfortunately, sometimes the analyses get way ahead of the data (one CEA was published on autologous bone marrow transplantation in metastatic breast cancer 8 years before a randomized trial showed no benefit).

#### 3. Do the effectiveness data reflect how the strategy will be used in the real world?

Even if the effectiveness data are from randomized trials, it's important to ask whether they really pertain to the population and setting in which the strategy is likely to be applied. Consider a CEA of carotid endarterectomy in asymptomatic patients with more than 70% stenosis. If the trial data represent the best surgical practice while broad implementation of the strategy would involve community providers, then effectiveness is being overestimated—as is cost-effectiveness. A similar problem may occur if the trials involve patient selection criteria that are not easily replicated in practice. A critical reader of CEAs should carefully consider the generalizability of the effectiveness data.

#### 4. Where do the cost data come from?

The basic question here is, "Was resource use modeled, or was it measured in real practice?" In modeling, investigators have to make assumptions about which services are likely to be utilized differently—thus driving the difference in cost. The measurement of resource use in practice has the advantage of capturing utilization that may not be anticipated by investigators (e.g., extra testing, extra visits, readmissions).

In either approach, there can be considerable debate about how to attach dollar amounts to utilization counts (debates that can get very tedious very quickly). Critical readers should look at the utilization counts themselves and have some confidence about the face validity of the dollars attached to them (probably the most practical standard being the Medicare fee schedule/allowed charges). If more utilization doesn't equal more money, something's wrong.

#### 5. Who's funding the CEA?

Unfortunately, funding sources seem to matter. There is now considerable evidence that researchers with ties to drug companies are indeed more likely to report favorable results than are researchers without such ties. Because they are so sensitive to both the choice of strategies and assumptions, CEAs are particularly susceptible to bias—intentional or not. Consequently, some journals have chosen not to publish industry-supported CEAs. For those that are published, readers must consider the conflict posed by funding from a manufacture of one of the analyzed strategies.

#### 6. Did we get anywhere?

Finally, readers may want to consider whether the entire exercise somehow helped them with a decision. Although some CEAs have extremely high CE ratios (i.e., > \$200,000 per quality-adjusted life-year—a poor value) and other have very low CE ratios (i.e., < \$10,000 per quality-adjusted life-year—a good value), most fall somewhere in the middle. Analyses with CE ratios of \$50,000 per quality-adjusted life-year may conclude with an assertion that the analyzed strategy is "cost-effective." Whether or not this helps anyone make a decision is hard to know.

A compendium of *ecp* primers from past issues can be viewed and/or requested at <http://www.acponline.org/journals/ecp/primers.htm>.

### Suggested Reading

Azimi NA, Welch HG. The effectiveness of cost-effectiveness analysis in containing costs. *J Gen Intern Med.* 1998;13:664-9.

Doubilet P, Weinstein MC, McNeil BJ. Use and misuse of the term "cost-effective" in medicine. *N Engl J Med.* 1986;314:253-6.

Drummond MF, Richardson WS, O'Brien BJ, Levine M, Heyland D. Users' guides to the medical literature. XIII. How to use an article on economic analysis of clinical practice. A. Are the results of the study valid? Evidence-Based Medicine Working Group. *JAMA.* 1997;277:1552-7.

Eddy DM. Cost-effectiveness analysis: a conversation with my father. *JAMA.* 1992;267:1669-75.

Eddy DM. Cost-effectiveness analysis: is it up to the task? *JAMA.* 1992;267:3342-48.

Eddy DM. Cost-effectiveness analysis: the inside story. *JAMA.* 1992;268:2575-82.

Eddy DM. Cost-effectiveness analysis: will it be accepted? *JAMA.* 1992;268:132-6.

Friedberg M, Saffran B, Stinson TJ, Nelson W, Bennett CL. Evaluation of conflict of interest in economic analyses of new drugs used in oncology. *JAMA.* 1999;282:1453-7.

Kassirer JP, Angell M. The Journal's policy on cost-effectiveness analyses. *N Engl J Med.* 1994;331:669-70.

O'Brien BJ, Heyland D, Richardson WS, Levine M, Drummond MF. Users' guides to the medical literature. XIII. How to use an article on economic analysis of clinical practice. B. What are the results and will they help me in caring for my patients? Evidence-Based Medicine Working Group. *JAMA.* 1997;277:1802-6.

Russell LB, Gold MR, Siegel JE, Daniels N, Weinstein MC, for the Panel on Cost-Effectiveness in Health and Medicine. The role of cost-effectiveness analysis in health and medicine. *JAMA.* 1996;276:1172-7.

Siegel JE, Weinstein MC, Russell LB, Gold MR, for the Panel on Cost-Effectiveness in Health and Medicine. Recommendations for reporting cost-effectiveness analyses. *JAMA.* 1996;276:1339-41.

Weinstein MC, Siegel JE, Gold MR, Kamlet MS, Russell LB, for the Panel on Cost-Effectiveness in Health and Medicine. Recommendations of the panel on cost-effectiveness in health and medicine. *JAMA.* 1996;276:1253-58.

Weinstein MC, Stasson WB. Foundations of cost-effectiveness analysis for health and medical practice. *N Engl J Med.* 1977;296:716-21.

## Primer on Interpreting Surveys

To answer their research questions, investigators often need to ask questions of others. These questions may revolve around how people feel, what people know, and what people think. Some examples are given in the following table.

GENERAL QUESTIONS	EXAMPLES
How do people feel?	How do patients with lung cancer feel after having chemotherapy? How do physicians react to having their decisions reviewed? How much do healthy women fear breast cancer?
What do people know?	What do patients know about the benefit of chemotherapy in lung cancer? What do physicians know about the evidence supporting certain therapies? What do women know about their risk for heart disease?
What do people think?	Do patients with lung cancer think they should be told the average survival benefit? Do physicians think that there is a better way to change their behavior? Do women think that they are getting too much or too little information?

To address these questions, investigators must systematically question a defined group of individuals—in other words, administer a survey. This can be done in person, by mail, by phone, or over the Internet. Because surveys are increasingly common in the medical literature, readers need to be able to critically evaluate the survey method. Two questions are fundamental: 1) Who do the respondents represent? 2) What do their answers mean?

### Who Do the Respondents Represent?

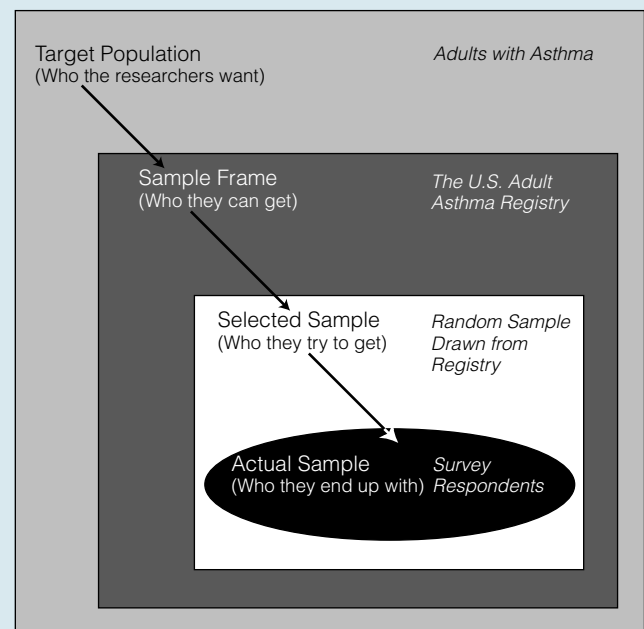
Like most types of research, surveys are useful only to the extent that they help us learn something about a defined population. The population we are interested in learning about is called the target population. Surveys are almost always based on a sample of the target population, and the respondents may not accurately represent this population.

Consider the following example. Suppose you are interested in how well adults with asthma are schooled in the use of spacers with inhalers. You question a sample of adults who are members of an asthma registry, and one third respond. You

are surprised by how educated most of them are about the technique. You conclude that there is little need for further education.

What's wrong with this conclusion? Patients in the registry may be more motivated than patients in general. Furthermore, patients who received the survey and did not know the answers to the questions might have decided not to complete it. Therefore, it is possible that your conclusion is wrong and that, in fact, most asthmatic persons do not understand the use of spacers.

To avoid this general problem, readers need to ask themselves how well the respondents represent the target population. As shown in the following figure, there are three basic steps of selection between the target population (about which the conclusion will be drawn) and the actual sample (where the data come from). The reduction at each step potentially threatens a conclusion about the target population.



### Target Population → Sample Frame

The sample frame is the portion of the target population that is accessible to researchers (e.g., persons who read newspapers, persons with phones). Often, the sample frame is some sort of list (e.g., a membership list). But individuals who are accessible may differ from those who are not. For example, persons with phones are different from persons without phones, and physicians who are members of professional organizations are different from those who are not. Readers should carefully judge how the sample frame might systematically differ from the target population.

## Sample Frame → Selected Sample

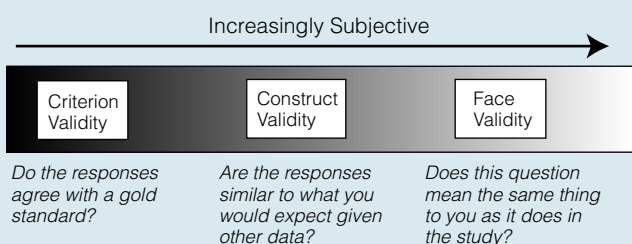
Although researchers may try to contact the entire sample frame, in many cases this would involve an unmanageable number of individuals. The selected sample is the portion of the sample frame that the researchers actually try to contact. If the selected sample is randomly selected from the sample frame, readers can be confident that this step does not seriously threaten generalizability. If it is selected by some other means, readers must be more circumspect. Suppose the selected sample is 100 patients who appear consecutively in an outpatient clinic (consecutive sample) or 100 persons who respond to a newspaper advertisement (convenience sample). Although both approaches are reasonable places to begin to learn about a topic, the first does not adequately represent patients coming to clinic (because it over-represents persons who visit the clinic frequently) and the second does not adequately represent persons who read newspapers.

## Selected Sample → Actual Sample

Not everyone who is contacted responds to a survey. The final sample is the portion of the selected sample that chooses to respond. However, the decision not to respond is usually not random—that is, respondents and nonrespondents usually differ. Patients who respond to questions about their disease may be more educated, have a smaller number of other problems, and care more about health. Physicians who respond to questions about guidelines may be more likely to believe that guidelines are important and more likely to be compliant. To judge these factors, readers need to consider the response rate. Whenever response rates are less than perfect (< 90%) and particularly when they are low (< 50%), readers should ask themselves how nonrespondents are likely to differ from respondents.

## What Do Their Answers Mean?

Having decided who the respondents represent, readers can proceed to making judgments about their responses. The real challenge is to think about validity: How well do the survey questions do their job? Validity is the degree to which a particular indicator measures what it is supposed to measure rather than reflecting some other phenomenon. Although there are numerous kinds of validity (and even more names for each kind), it may be more useful for readers to consider validity as a spectrum, as in the following illustration.



### Criterion Validity

At one extreme, readers can determine the extent to which researchers have compared the performance of their question

with an external gold standard. Examples of criterion validity include comparing reported age with birth certificates, reported weight with measured weight, and reported eyesight with visual acuity. Although readers may be much more confident about a question that has been validated against an explicit criterion, they must also ask whether it may have been more accurate to simply apply the gold standard (e.g., why ask about weight when you can measure it?). Unfortunately, there is no criterion for many important questions (e.g., questions about what people think).

### Face Validity

At the other extreme, readers need to consider for themselves whether the questions seem appropriate and reasonably complete “on the face of it.” To really judge face validity, readers should look (and journals should publish) the exact language used in the question. Face validity has the disadvantage of being entirely subjective. At the same time, it may be the only type of validity that can be applied to the important subjective questions that survey researchers are trying to answer.

### Construct Validity

Construct validity is somewhere between criterion validity and face validity. When the “gold standard” is not very objective but other data are available with which to judge a question’s performance, we are in the realm of construct validity. The basic idea behind construct validity is that if your measurement does what you think it does, it should behave in certain ways. For example, the level of self-reported pain would be expected to decrease when respondents are given morphine. Wherever possible, readers should look for evidence that the pattern of responses is generally what would be expected given other data.

### Interpreting Scores

It is increasingly common to see the answers for several questions aggregated into a single score (“The mean PDQ score for dentists was 2.5 points higher than for lawyers;  $P = 0.03$ ”). If possible, readers should try to move beyond the score to consider the validity of individual questions. But because use of scores is increasing, readers also need to seek some grounding about what the scores mean (“Is 2.5 big or little?”). Sometimes this grounding can be achieved by knowing the mean score for groups with which one is familiar or by knowing how much a score changes after a familiar event. Knowing that the development of a new chronic disease translates to approximately a 5-point drop in the Physical Component Summary score of the SF-36, for example, helps give a sense for this measure of health status.

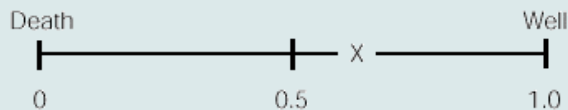
## Conclusions

Survey research is an important way of learning what our patients understand and what they want. At the same time, it is often cluttered with unnecessary complexity and jargon. More important, false conclusions are a constant possibility. Simply figuring out what questions were asked and who the respondents were will go a long way toward avoiding these problems.

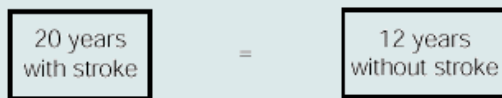
## Primer on Utilities

Utilities are numerical expressions of patient preferences for a particular state of health. Although utilities and measures of functional status both reflect quality of life, utilities describe how patients feel about or value living with a given clinical condition, and measures of functional status generally reflect the limitations experienced by patients with a clinical condition (e.g., New York Heart Association class for congestive heart failure). Utilities are typically assessed on a scale from 0 (death or worst health imaginable) to 1 (best health).

### Visual Analogue Scale



### Time Trade-off



### Standard Gamble

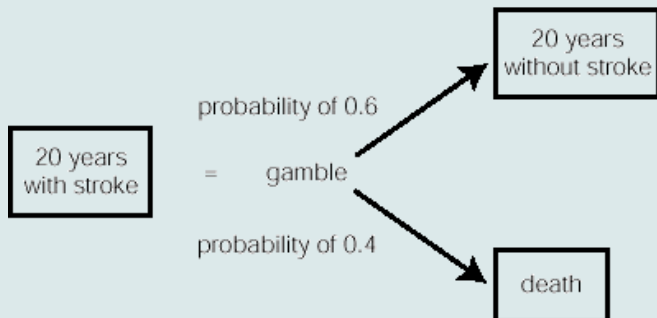


FIGURE. Three ways to measure or express a utility of 0.6 for disabling stroke.

Patient utilities may be measured by using a variety of techniques (Figure). With the simplest approach, the visual analogue scale, patients simply mark an "X" on a continuous

scale between 0 and 1. More commonly, utilities are elicited by asking patients to make a series of choices to identify at what point they are indifferent about the choice between two options. There are two commonly used iterative approaches to assessing utilities. With the time trade-off method, for example, patients might be asked whether they would prefer to live 10 years in good health or 20 years with a disabling stroke. If they chose the latter, the choice might be modified to 15 years in good health or 20 years living with a disabling stroke. This iterative process would continue until a patient was indifferent about the choice between the two options—for example, that living 12 years in good health was equivalent to living 20 years with a disabling stroke. In this case, the utility for stroke is the ratio of the two values:  $12/20=0.6$  (Figure). With the standard gamble method, a patient is instead asked to choose between life with a specific condition and a gamble with variable probabilities of life without the condition and death.

Average utilities for a wide variety of clinical conditions or symptoms may be obtained from the literature. One often-used catalogue is the Beaver Dam study.<sup>1</sup> This population-based study describes utilities (obtained by two different methods) for patients with a variety of common clinical conditions, such as severe back pain (0.87), insulin-dependent diabetes (0.72), and cataract (0.94).

One familiar application of utilities is the quality-adjusted life-year (QALY). To calculate QALYs, time spent in a particular outcome state is multiplied by the utility for life in that state. For example, 10 years after a disabling stroke (utility of 0.6) is equivalent to 6.0 QALYs ( $10 \times 0.6 = 6.0$  QALYs). This aggregate measure is frequently used in decision analysis and cost-effectiveness analysis to compare the relative value of clinical interventions.

### Reference

1. Fryback DG, Dasbach EJ, Klein R, et al. The Beaver Dam Health Outcomes Study: initial catalog of health-state quality factors. *Med Decis Making*. 1993;13:89-102.

## A Primer on Scores: What Counts?

To judge the effects of clinical interventions, researchers look for changes in certain key variables—better known as outcome measures. Some of the most familiar (and most important) outcome measures are dichotomous variables (so-called “0,1 variables”): They either happen or they don’t. Examples include heart attacks, strokes, and death. Other outcome measures can take on many values. Physiologic and laboratory measurements fall into this category (such as blood pressure, serum sodium levels, and CD4 counts), as do various functional status and symptom scales (such as the Glasgow Coma Scale to classify level of consciousness and visual analog scales to classify level of pain).

Over the past two decades, a new type of outcome measure has been increasingly used in clinical research: scores. A score is a composite measure—in other words, it is derived from several individual variables. A score may be the composite of multiple dichotomous variables, multiple physiologic and laboratory measurements, multiple scales, or any combination thereof. Scores are used primarily to measure multiattribute patient function (e.g., Mini-Mental Status Score is a metric for classifying the combined functions of orientation, computational ability, and short-term memory) or to predict risk for various outcomes (e.g., heart attack, breast cancer, or death).

Because they may summarize several different variables (which may have various weights), it can be difficult to know what a score really means. If the topic is of interest and primary outcome is a score, critical readers should seek answers to the following questions (Table 1). (If you can’t answer these questions, it’s tough to know what counts as an important effect.)

**TABLE 1**  
**QUESTIONS TO ANSWER TO UNDERSTAND A SCORE\***

QUESTION	ANSWER
What’s being measured?	
Which end is up?	
What’s possible?	
What are some benchmarks?	
What matters?	

\*Finding the answers can be challenging. One excellent resource for understanding functional health scores is McDowell I, Newell C. *Measuring Health, 2nd ed.* Oxford: Oxford Univ Pr; 1996.

### *What’s being measured?*

The first step is to try to get a handle on the construct. This can be harder than you think. Like so many things in medicine, scores often go by their acronym (and even when you know what the acronym stands for, you may not be that much closer to the construct). Consider the following examples. PCS stands for physical component summary; it is an overall measure of physical function assessed by self-report (part of the Medical Outcomes Study SF-36). APACHE II stands for Acute Physiologic and Chronic Health Evaluation (second version); it is a prognostic measure for intensive care unit patients that is used to predict inpatient mortality.

### *Which end is up?*

Sometimes it’s hard to know whether a higher score is a good thing or a bad thing. A high PCS score, for example, is good. A high APACHE II score, on the other hand, most definitely is not.

### *What’s possible?*

Knowing the range of possible values is the next step for getting a feel for the results. Some scores, such as the PCS score, range from 0 to 100. But many do not (APACHE II ranges from 0 to 71).

### *What are some benchmarks?*

The reader needs context—some grounding on what an expected score would be for a defined set of individuals. Published norms are available for the PCS score.<sup>1</sup> For example, in the general U.S. population, the average PCS score for men over 65 years is 42. A healthy 40-year-old will have an APACHE II score of 0.

### *What matters?*

Finally, the reader needs help to make judgments about what constitutes an important change. In other words, a reader needs a clinical correlation. A 5-point decrease in the PCS score, for example, is equivalent to developing a new chronic disease like congestive heart failure. Of course, the information is not as precise as we would like (the severity of congestive heart failure varies from person to person, as does its impact), but it’s a lot better than nothing. A change in APACHE II from 12 to 24 is associated with an absolute increase in inpatient mortality of 30% (from approximately 10% to over 40%).

To make sense of scores, readers should try to answer the preceding questions. Unfortunately, authors often fail to provide the needed information. In these cases, if readers want to really understand what a score means, they must do the hard work themselves.

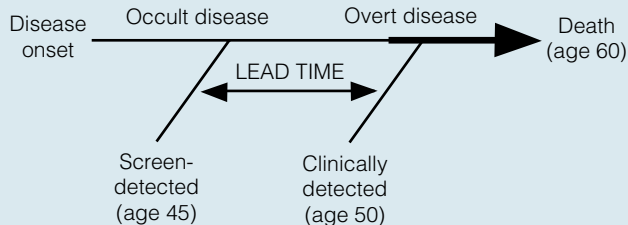
### **Reference**

1. SF-36 Physical and Mental Health Summary Scales: A User’s Manual. Boston: The Health Institute, New England Medical Center; 1994.

## Primer on Lead-Time, Length, and Overdiagnosis Biases

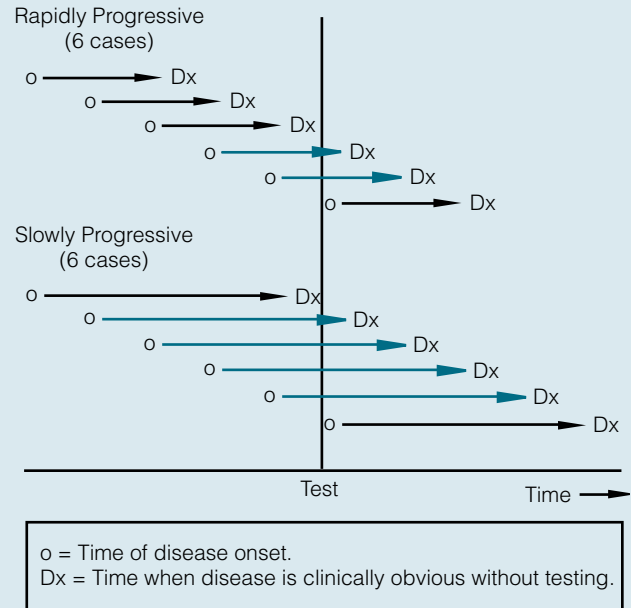
The apparent effects of early diagnosis and intervention (measured in terms of how screening-detected cases compare with cases detected by signs and symptoms) are always more favorable than the real effects (measured in terms of how a population that is screened compares with a population that is not). The comparison between screening-detected cases and others overestimates benefit because the former consists of cases that were diagnosed earlier, progress more slowly, and may never become clinically relevant. This comparison, therefore, is said to be *biased*. In fact, three biases exist that inflate the survival of screen-detected cases.

1. **Lead-time bias: Overestimation of survival duration among screen-detected cases (relative to those detected by signs and symptoms) when survival is measured from diagnosis.** In the figure below (representing one patient), the patient survives for 10 years after clinical diagnosis and survives for 15 years after the screening-detected diagnosis. However, this simply reflects earlier diagnosis because the overall survival time of the patient is unchanged.



2. **Length bias: Overestimation of survival duration among screening-detected cases caused by the relative excess of slowly progressing cases.** These cases are disproportionately identified by screening because the probability of detection is directly proportional to the length of time during which they are detectable (and thereby inversely proportional to the pro-

gression). In the following figure (representing 12 patients), 2 of 6 rapidly progressive cases are detected, whereas 4 of 6 slowly progressive cases are detected.



3. **Overdiagnosis bias: Overestimation of survival duration among screen-detected cases caused by inclusion of *pseudodisease*—subclinical disease that would not become overt before the patient dies of other causes.** Some researchers further divide pseudodisease into two categories: one in which the disease does not progress (type I) and another in which the disease does progress—but so slowly that it never becomes clinically evident to the patient (type II). Inclusion of either type as being a “case” of disease improves apparent outcomes of screening-detected cases.

## Primer on Dissecting a Medical Imperative

Clinicians often face medical imperatives, which are broad statements that endorse a course of action. Consider two familiar medical imperatives: invest in patient safety and screen for cancer. Supporting these imperatives are the assertions that eliminating mistakes and early cancer detection will save lives.

Medical imperatives are rarely the result of a single study. Instead, they are generally the product of a complex mixture of observation, reasoning, and belief. Because the actions they engender may be beneficial, distracting, or possibly even be harmful, critical readers will want to carefully consider the line of reasoning on which they are based. Several steps may be useful in this regard.

### Diagram the Line of Reasoning

Diagramming the argument that supports an imperative provides the structure necessary to carefully consider the issue. Figure 1 is a prototype for the line of reasoning for each of the above examples (other constructions are, of course, possible).

### Understand the Vocabulary

The process of depicting the argument also helps to identify critical issues of definition (e.g., What constitutes an error? What

constitutes cancer?) that may have important implications when the imperative is put into action (e.g., Do doctors agree on what an error is? Do pathologists agree on who has early cancer?). Carefully understanding the vocabulary may also help identify subtle changes in words (e.g., from preventable adverse event to error) that may have tremendous influence on public policy.

### Distinguish between Observation and Inference

Once an argument is diagrammed, each element should be considered in terms of its source. Is it the product of an observation or the result of an inference? Generally, the observations appear earlier in the line of argument.

### Critically Examine the Observations

The observations are typically the result of published findings and should be subject to the same scrutiny given any important finding (e.g., Is it relevant? Is it valid? Is it generalizable?).

### Look Out for Leaps of Faith

Next, consider the inferences carefully. Some may be cautious and conservative, others may be reckless. The most common problem is to confuse association and causality (e.g., "Because people who

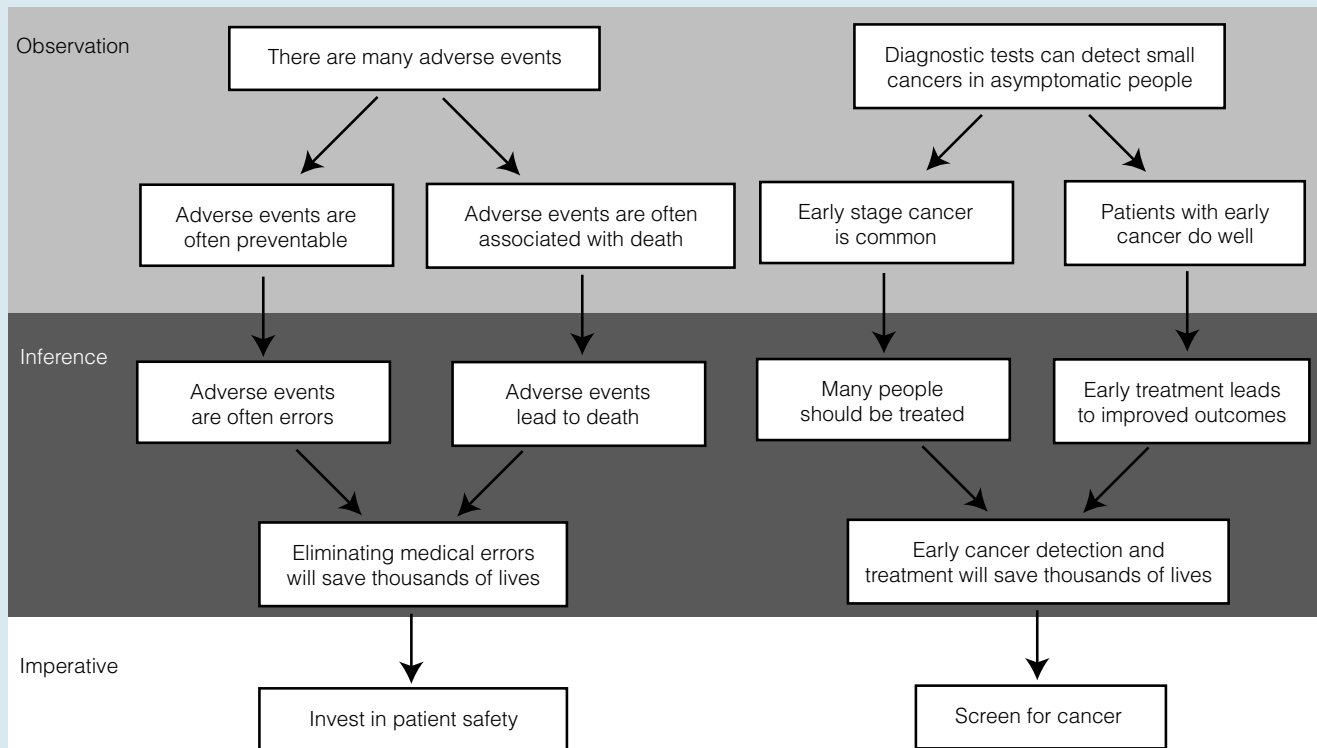


FIGURE 1. Lines of reasoning underlying two imperatives.

die in the hospital often experience adverse events, preventing adverse events will save lives” or “Because patients with early disease do well, early treatment will improve outcomes”).

### **Ask about Vested Interest**

How impartial is the person (or group) promoting the imperative? Obviously, some degree of intellectual interest is expected. But the presence of strong professional and/or financial interests may unduly influence the call for action (e.g., safety consultants call for safety initiatives, mammographers calling for mammography).

### **Consider Unintended Effects**

Finally, think hard about the net effects (intended and unintended) of the proposed course of action. Even the simplest action

can have unintended effects. For example, cancer screening may help some people avoid late-stage disease, yet lead others to be treated unnecessarily (e.g., those with nonprogressive cancer). And all actions have opportunity costs. For example, dollars devoted to nurse clinicians to improve patient safety are dollars taken from something else. If that something is routine hospital nursing services, the net effect may be to diminish patient safety. Just because net effects are difficult to predict, it doesn't mean they can be ignored.

It's important to think about medical imperatives carefully. When you do so, you will probably find that most are oversimplifications. Unfortunately, the world is more complex than any of us would like. Most imperatives are probably neither right nor wrong—instead, there are settings where they are useful and others where they are not.

*A compendium of **ecp** primers from past issues can be viewed and/or requested at <http://www.acponline.org/journals/ecp/primers.htm>.*

## A Primer on HEDIS

Although many people talk about report cards for medical care, there are few working examples. The most prominent is the Health Plan Employer Data and Information Set, better known as HEDIS. Used by over 400 health plans, HEDIS is a set of standardized performance measures intended to help purchasers and patients compare health plans in terms of quality (instead of simply comparing costs).

HEDIS is perhaps best thought of as a standardized test for health plans. As in most standardized tests, different sections test different domains (e.g., mathematics, language skills). Each domain contains a series of performance measures (e.g., individual questions). Table 1 shows the seven HEDIS domains and selected performance measures.

**TABLE 1**  
**HEDIS Domains\***

DOMAIN	SELECTED PERFORMANCE MEASURES
Effectiveness of care	See Tables 2 and 3
Access and availability of care	Proportion of enrollees with preventive/ambulatory health visits during the reporting year (calculated separately for children and adults) Number of providers (primary, behavioral health, obstetric and prenatal, and dental) Availability of language interpretation services
Satisfaction with experience of care	Member satisfaction
Health plan stability	Disenrollment Provider turnover Indicators of financial stability (e.g., revenue, loss, reserves held by plan)
Use of services	Visits (prenatal care, well-child, adolescent well-care, other ambulatory care) Frequency of selected procedures Cesarean section rate Vaginal birth after cesarean rate Inpatient utilization (acute care, maternity care, newborns, mental health, chemical dependency) Outpatient drug utilization
Cost of care	Actual expense per member per month High-occurrence/high-cost DRGs (e.g., stroke, TIA, pneumonia, asthma, COPD, chest pain, angina pectoris, heart failure and shock, major joint replacement)
Health plan descriptive information	Total enrollment and enrollment by payer Provider characteristics (board certification, residency completion, compensation) Report of plan affiliations with public health, community-based and school-based agencies Cultural diversity of Medicaid membership

\*COPD = chronic obstructive pulmonary disease; DRG = diagnosis-related group; TIA = transient ischemic attack.

HEDIS measures of greatest interest to clinicians are in the effectiveness-of-care domain. Table 2 lists the performance measures, describes how each is calculated, and reports the most recent averages available for the Alliance of Community Health Plans and the national average (representing all participating plans). In each case, a higher proportion is presumed to represent better care. Some patients, however, may have an informed preference to forgo some of these services, such as certain immunizations (see the article by Mehl in this issue).

The individual performance measures have evolved over time. When HEDIS was initiated in 1991, the effectiveness measures focused on vaccination and screening rates. Measures were added

subsequently to reflect treatment quality in diabetic and post-myocardial infarction patients. New measures to examine care of patients with hypertension, asthma, chlamydia, and menopause have been proposed for the next version of HEDIS (Table 3).

As HEDIS performance measures become more complex, so do the questions about measurement methods (e.g., Does a blood pressure of 145/95 mm Hg require control? What constitutes a sufficient discussion of treatment options?).

HEDIS is managed by the National Committee for Quality Assurance (NCQA). NCQA is encouraging the broad use of HEDIS data by employers, consumers, and other health care professionals to compare health plans. Further information can be found at [www.ncqa.org](http://www.ncqa.org).

TABLE 2

## Current Performance Measures in the Effectiveness-of-Care Domain\*

PERFORMANCE MEASURE	NUMERATOR	DENOMINATOR	1997 ACHP AVERAGE	1997 NATIONAL AVERAGE
Childhood immunization rate	DPT, polio, MMR, hepatitis B, HIB	2-yr-olds	78%	65%
Adolescent immunization rate	2nd MMR, hepatitis B, chicken pox	13-yr-olds	12%	8%
Advice to quit smoking	Received advice to quit	Adults $\geq$ 18 yr who are current smokers	69%	64%
Breast cancer screening rate	One or more mammo-grams in the past 2 years	Women aged 52–69 yr	77%	71%
Cervical cancer screening	One or more Pap tests in the past 3 years	Women aged 21–64 yr	77%	71%
Rate of prenatal care in the first trimester	Prenatal care visit between 176 and 280 days before delivery	Women who delivered live babies	88%	83%
Check-ups after delivery	Postpartum visit between 21 and 56 days after delivery	Women who delivered live babies	73%	66%
$\beta$ -blocker treatment rate	$\beta$ -blocker dispensed within 7 days after AMI discharge	Adults $\geq$ 35 yr admitted with a diagnosis of AMI	79%	74%
Diabetic retinal examination rate	Retinal examination by an eye care professional	Adults $\geq$ 31 yr who have diabetes	53%	39%
Rate of follow-up after hospitalization for mental illness	Visit with mental health provider within 30 days of discharge	Individuals $\geq$ 6 yr admitted with a mental health diagnosis	77%	67%

\*ACHP = Alliance of Community Health Plans; AMI = acute myocardial infarction; DPT = diphtheria, pertussis, tetanus; HIB = Haemophilus influenzae type B; MMR = measles, mumps, and rubella; Pap = Papanicolaou.

TABLE 3

## New Effectiveness-of-Care Performance Measures for HEDIS 2000

PERFORMANCE MEASURE	NUMERATOR	DENOMINATOR
Controlling high blood pressure	Blood pressure controlled to below 140/90 mm Hg	Enrollees with high blood pressure
Appropriate medications for people with asthma	Received medications for long-term control (e.g., inhaled corticosteroids)	Enrollees with chronic asthma
Chlamydia screening	Tested for chlamydia	Sexually active women aged 15–25 yr
Management of menopause*	Breadth, depth, and personalization of menopause counseling	Menopausal women

\*This measure encourages plans to discuss with women the pros and cons of various treatment options, such as hormone replacement therapy, so that they can make more informed choices.

# Primer on Geographic Variation in Health Care

Although regional variation in health care has long been recognized,<sup>1</sup> studies describing variation in intervention rates across geographic areas continue to appear regularly in medical journals. This primer is intended to help readers make sense of reports about geographic variation. We focus on two basic questions: 1) How much variation is there? 2) What causes variation?

## How Much Variation Is There?

Regional rates of medical interventions always vary. Chance alone creates some degree of variation in intervention rates, particularly when many geographic areas are compared. Although debate remains about which method is best, a variety of statistical approaches can be used to evaluate the role of chance in studies of geographic variation.<sup>2</sup>

In most studies, however, geographic variation in intervention rates is not due to chance alone (i.e., it is statistically significant). So readers must consider the “clinical significance” of observed variations: How much variation is there? Many studies simply report the extremal range (ratio of highest to lowest rates) to reflect the magnitude of variation (e.g., “rates of carotid endarterectomy varied 7-fold, from 1.1 to 7.6 per 1000 enrollees”).<sup>3</sup> However, this measure can be misleading because procedures performed infrequently generally appear more variable than more common procedures. The extremal range also reflects rates only in high and low outlier regions, thus ignoring practice patterns in all other regions.

To compare procedures reliably, variation measures should be standardized (i.e., on the same scale). One approach is to divide observed procedure rates in each region by the overall

average. As illustrated in Figure 1, plotting standardized rates is useful for comparing the “variation profiles” of different procedures.<sup>4</sup> Some procedures, such as hip fracture repair and colectomy for colon cancer, vary little—regional rates cluster near the national average. In contrast, radical prostatectomy and back surgery vary markedly—their variation patterns are scattered diffusely. Peripheral arterial angioplasty varies even more than these high-variation benchmarks.<sup>5</sup>

## What Causes Variation?

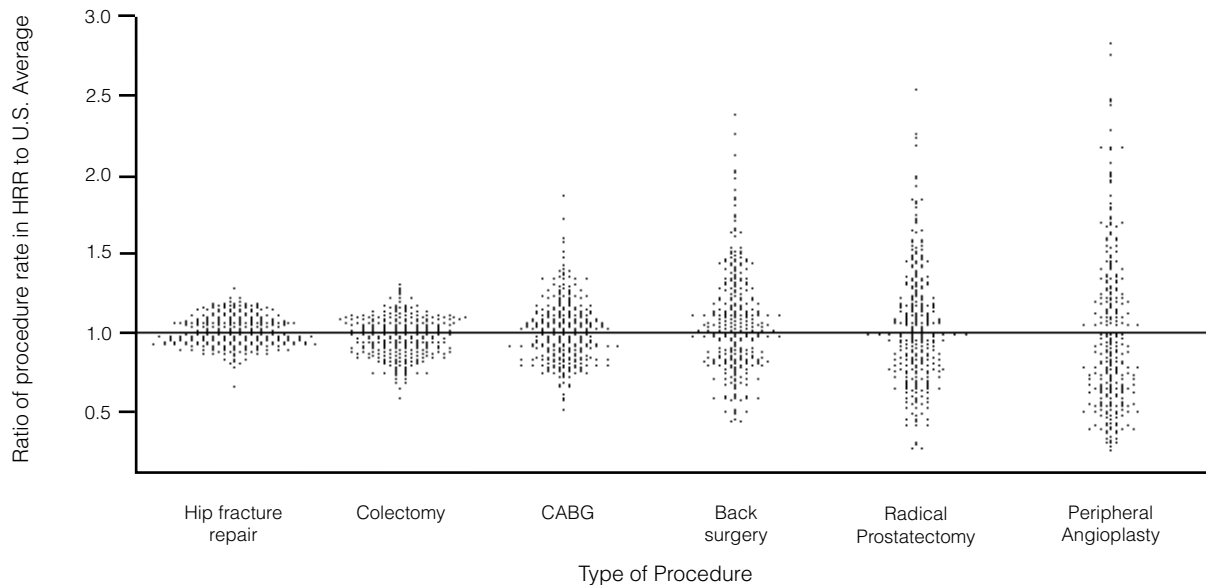
Considering the entire sequence of steps by which a patient ultimately gets to surgery (or any medical intervention) is a useful way to understand the potential explanations for geographic variation (Figure 2).

### Prevalence of Disease

Procedure rates may vary because of underlying differences in disease prevalence across regions. For example, generally higher rates of cardiovascular interventions in the southeastern United States may be in part related to a higher prevalence of cigarette smoking and other risk factors in that region.

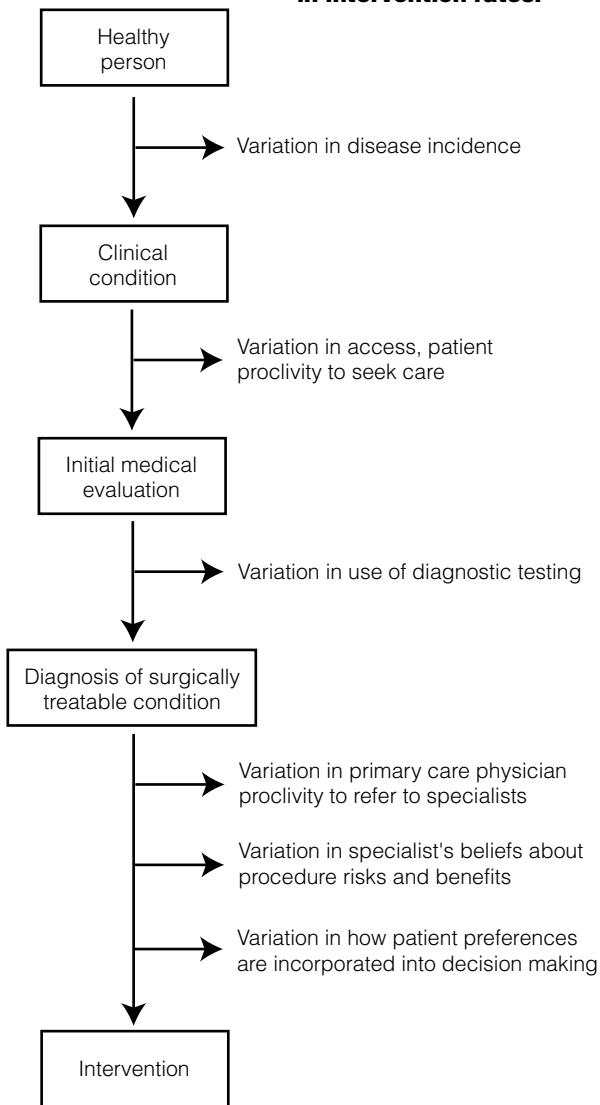
### Access to Care

To receive a procedure, patients must first get into the medical system. Procedure rates may vary if there are regional differences in access (e.g., related to socioeconomic status, insurance) or patient proclivity to seek medical care (e.g., related to race/culture).



**FIGURE 1. Variation profiles of six common procedures.** Data for peripheral angioplasty from Axelrod and colleagues.<sup>3</sup> Other figures derived from 1995–6 national Medicare data from the *Dartmouth Atlas of Health Care*.<sup>5</sup> CABG = coronary artery bypass grafting; HRR = hospital referral region.

**Potential reasons for geographic variation in intervention rates:**



**FIGURE 2. Process by which a healthy person becomes a patient and ultimately receives a medical intervention and potential reasons for geographic variation in intervention rates.**

**Decision To Test**

Many surgically treatable conditions are identified primarily by diagnostic tests (e.g., prostate-specific antigen testing, coronary angiography). Thus, surgery rates may vary because of regional variation in the use of diagnostic testing. For example, regional rates of carotid endarterectomy have been shown to be highly correlated with rates of carotid ultrasonography.<sup>3</sup>

**Decision To Treat**

Finally, it is important to consider how treatment decisions are made, particularly in instances where treatment is not constrained to a single therapeutic option. Several components of this decision process may contribute to regional variation in intervention rates. Primary care physicians may vary in their propensity to refer patients to specialists (and delegate decision making to them). Specialists may vary in their beliefs about the risks and benefits of a given procedure, and thus vary in the recommendations they give patients. Finally, there may be regional variation in the degree to which individual patient preferences are incorporated into clinical decisions.

**Conclusion**

Differences in the degree to which procedures vary can be explained in the context of these components of decision making. Consider hip fracture repair, a low-variation procedure. Hip fracture prevalence does not vary geographically—all patients seek care, the diagnosis is usually made without discretionary testing, and decisions about treatment are constrained to a single option (surgery). In contrast, regional rates of radical prostatectomy vary widely. This is not surprising: Prostate cancer prevalence varies widely (likely due to variation in testing), and there is wide disagreement among both primary care physicians and specialists about the risks and benefits of several different treatment options.

Geographic variation studies often identify unrecognized problems in clinical decision making. These studies stimulate us to ask, but cannot answer, the question, “Which rate is right?” Research aimed at better understanding of clinical effectiveness, patient preferences, and economic implications is necessary for addressing this basic question.

**References**

1. Wennberg JE, Gittelsohn A. Small area variation in health care delivery. *Science*. 1973;182:1102-8.
2. Diehr P, Cain K, Connell F, Volinn E. What is too much variation? The null hypothesis in small area analysis. *Health Serv Res*. 1990;24:741-71.
3. Wennberg JE, Cooper MM. Practice variations and the quality of surgical care for common conditions. In: 1999 Dartmouth Atlas of Health Care. Chicago: American Hospital Publishing; 1999.
4. Birkmeyer JD, Sharp SM, Finlayson SRG, Fisher ES, Wennberg JE. Variation profiles of common surgical procedures. *Surgery* 1998;124:917-23.
5. Axelrod DA, Fendrick AM, Wennberg DE, Birkmeyer JD, Siewers AE. Cardiologists performing peripheral angioplasties: impact on utilization. *Eff Clin Pract*. 2001;4:191-8.

**Acknowledgment**

This primer was contributed by John D. Birkmeyer, MD, Dartmouth Medical School, Hanover, New Hampshire.