Stat.BD

CLASS NOTES • CHAPTER 1

What Statistics Does in Medical Research

Comparison, inference, and how to read a result

Applied Biostatistics & Research Methodology for Clinical Research
Stat.BD mentorship programme

Mentor: Muhtasim Munif Fahim, MSc
Department of Statistics & Data Science, University of Rajshahi

Chapter 1

What Statistics Does in Medical Research

Why clinical experience cannot settle a question about treatment, what a research paper is built to do about that, and how to read the three separate claims hidden inside a single sentence of results.

Medicine accumulated knowledge for a long time before it accumulated method. The change, when it came, was not the arrival of better doctors or better drugs. It was the arrival of a way to settle disagreements that did not depend on who was arguing. This chapter is about that way of settling things, what it can do, and the large class of problems it cannot touch.

Order matters here. The chapter begins with the problem that makes research necessary at all, which is a problem about knowledge rather than about numbers. Only then does it turn to the machinery: how papers are organised, how questions are framed, what separates a fact about the people in a study from a claim about everyone else, and how to take a result sentence apart. Formulas appear from Section 6 onwards, always stated rather than derived, and always followed by the same numbers a published trial reported.

What this chapter establishes

1. Why a comparison group is not a formality but the entire basis of a claim that a treatment works.
2. How a medical paper is organised, what each section promises, and the order in which its sections are best read.
3. How to convert a clinical problem into a question a study could actually answer.
4. How to tell description from inference in any paper, which is the distinction almost every argument about evidence turns on.
5. How to read magnitude, precision and evidence out of a single reported result, and how to convert between absolute and relative expressions of the same finding.

No mathematical background is assumed. The chapter uses arithmetic and no algebra beyond substitution.

Notation

Symbol	Meaning
n	the number of people in a sample, or in one group of it
π	the true proportion in the population, which is unknown
$\hat{p}$	the proportion observed in the sample, which estimates π
$\hat{p}_1,\ \hat{p}_0$	that proportion in the treated and the comparison group
μ	the true mean in the population, which is unknown
$\bar{x}$	the mean observed in the sample, which estimates μ
$Y_i(1),\ Y_i(0)$	what would happen to patient i with, and without, the treatment
RD	risk difference, an absolute measure
RR	relative risk, a ratio
OR	odds ratio, a ratio of odds rather than of risks
NNT	number needed to treat
CI	confidence interval

A hat over a symbol means an estimate calculated from data. A symbol without one refers to the underlying truth, which no study observes directly. That distinction is the whole of Section 4 and it is worth noticing early.

Contents

1 The problem of knowing

Consider a question put to a doctor with ten years of clinical practice behind them: given all those patients, how is it known which treatments work? The question sounds rhetorical and is not. The answer is less obvious than it appears, and most of research methodology follows from taking it seriously.

There are three grounds on which a clinician comes to believe that something is true, and they are not equally reliable.

1.1 Authority

A professor said so, a textbook says so, a guideline says so. Most of what any clinician knows arrives this way, and it has to, because nobody can personally verify the whole of medicine. A working doctor who insisted on first-hand evidence for every claim would never get through a morning clinic.

Authority fails in one specific way. It fails when the chain has no end, when the professor was repeating their own teacher, who was repeating theirs. Somewhere along that chain somebody has to have actually looked, and the only question that matters is whether they did. Authority is therefore not an alternative to evidence. It is a delivery mechanism for it, and like any delivery mechanism it can be carrying nothing.

1.2 Experience

This is the ground that feels most solid and is the most treacherous.

Suppose a clinician sees a patient with a sore throat, prescribes an antibiotic, and finds the patient well four days later. What did the antibiotic do? The honest answer is that nothing in that encounter can say. Most sore throats settle in about that time without any treatment at all. The improvement was real and observed; the attribution of it to the antibiotic was supplied by the observer. Nor can the attribution ever be checked, because there was only one of that patient and they have already been treated.

That unobserved version, the same patient at the same moment left alone, is the **counterfactual**. Its permanent absence is the reason personal experience cannot settle a question about treatment, however much experience there is. From the bedside, improvement under treatment and spontaneous recovery look identical. So do a drug that works and a disease that was going to remit anyway. Ten years of practice produces ten years of observations with the comparison missing from every single one.

Counterfactual
What would have happened to the same patient under the other treatment. It is never observed, in any study, ever.

Definition 1.1 • The counterfactual, stated formally

For a patient i, write $Y_i(1)$ for the outcome that would occur if that patient received the treatment, and $Y_i(0)$ for the outcome that would occur if the same patient did not. The effect of the treatment on that patient is the difference between them:

$$\tau_i \;=\; Y_i(1) - Y_i(0). \tag{1.1}$$

Definition 1.1 • *continued*

Exactly one of $Y_i(1)$ and $Y_i(0)$ is ever observed. The other is the counterfactual, and it is missing by definition rather than by oversight.

$Y_i(1)$ the outcome for patient i if treated
$Y_i(0)$ the outcome for the same patient i if not treated
τ_i the treatment effect for that one patient

Equation (1.1) is worth pausing on, because it is a formula that can never be evaluated. Both terms refer to the same patient at the same moment, and a patient is either treated or not. This is sometimes called the fundamental problem of causal inference, and the name is fair: it is not a limitation of current technique but a feature of the world.

1.3 From one patient to a group

If τ_i cannot be recovered for any individual, something weaker has to be settled for instead. What research estimates is not the effect on a particular patient but the average effect across a group:

$$\tau = \mathrm{E}[Y(1)] - \mathrm{E}[Y(0)], \tag{1.2}$$

the average outcome if everyone in some population were treated, minus the average outcome if nobody were.

$\mathrm{E}[\cdot]$ the average value across the population
τ the average treatment effect

This is a genuine retreat and it should be recognised as one. Equation (1.2) says nothing about whether a particular patient will benefit. It describes what happens on average, which is why a treatment with a real average benefit still fails some of the people who receive it.

Intuition

The trade in (1.2) is this. The individual effect in (1.1) is the quantity everyone actually wants and nobody can have. The average effect is a weaker quantity that can be estimated, because although no single patient is observed both ways, a group of treated patients and a group of untreated patients can both be observed once. Every study design in this course is a way of making the second group a believable stand-in for what the first group would have done untreated.

Why this governs everything that follows

A control group is a stand-in for the counterfactual. A comparison ward is a stand-in. Matching is a stand-in. Randomisation is the most convincing stand-in anyone has yet devised, because it makes the two groups comparable in respects nobody thought to measure as well as in the ones they did. Seen this way, the long list of study designs in Phase II stops being a list to memorise and becomes a series of answers to a single problem.

1.4 Evidence

Evidence is the third ground: a record of what happened when a comparison was actually made. It costs time, money, effort and a willingness to be proved wrong, which is why there is less of it than there should be.

Evidence is not a higher form of knowing that supersedes the other two. Clinical judgement decides which questions are worth asking, and no amount of method substitutes for it. The claim here is narrower. Judgement cannot, on its own, settle whether a treatment works, because the evidence that would settle it was never available at the bedside. What evidence adds is that the comparison was made deliberately enough for somebody who was not there to check it.

1.5 A hundred years of learning one thing

This history is short, and it repeats. In 1847 Ignaz Semmelweis worked in a Vienna hospital with two maternity clinics, one staffed by doctors and medical students and one by midwives. Mothers begged to be admitted to the second, and they were right to: over the preceding seven years, deaths ran at 98.4 per thousand births in the doctors' clinic against 36.2 in the midwives'. He noticed that the students came to the delivery room straight from the autopsy room, required them to wash in chlorinated lime, and watched monthly mortality in his clinic fall to around two per cent.

What Semmelweis contributed was not handwashing. It was the shape of the argument. Two groups, alike in city, year, hospital and patients, differing in one identifiable respect, with the outcome counted in both. That is a controlled comparison, and it arrived a century before anyone had the vocabulary for it.

Seven years later John Snow plotted cholera deaths on a map of Soho, one stacked bar per death at the address where it occurred, and the deaths crowded around a single water pump. Four years after that Florence Nightingale drew army mortality in a form that a Member of Parliament could not look away from, and the army medical service was reformed. In 1948 the Medical Research Council allocated scarce streptomycin by a schedule of random numbers kept in sealed envelopes, so that the admitting doctor could not know, and therefore could not influence, which arm the next patient entered. In 1950 Doll and Hill, unable to randomise anybody to smoke, built a design that starts from the disease and looks backward at the exposure.

Year	Who	What was added to the method
1747	Lind	Similar patients, same conditions, compared at the same time
1830s	Louis	Counting outcomes instead of asserting them
1847	Semmelweis	Two groups alike in everything but one factor
1854	Snow	Data as evidence, and a natural experiment
1858	Nightingale	Showing evidence so that it cannot be ignored
1948	MRC	Random allocation, with the schedule concealed
1950	Doll and Hill	A design for exposures that cannot be assigned
1996	CONSORT	Honest reporting made a condition of publishing

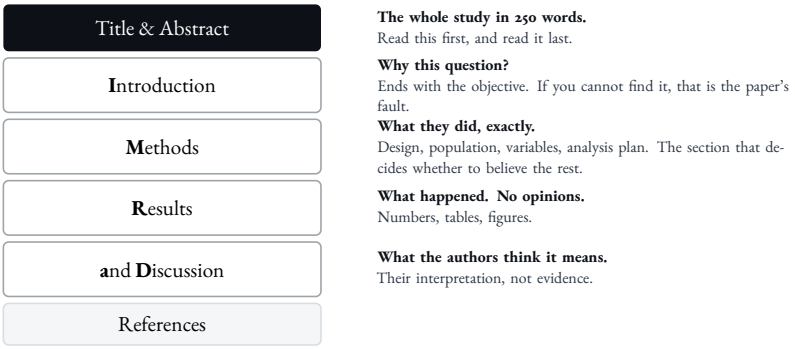
Notice what is absent from all of it. Not one of these people began with statistics. Every one began with something they had noticed and could not let go of: mothers dying in the

wrong ward, deaths clustering on a map, soldiers dying of the wrong thing. The method came afterwards, as the machinery for settling the argument. That order is why this course spends ten sessions on questions and design before it teaches a single test.

2 The anatomy of a research paper

Every modern medical paper has the same skeleton, and knowing it turns reading from a slog into a search. The structure is called **IMRaD**, after its four main sections. It is not a law of nature. It spread through medical journals during the twentieth century and only became near-universal in the 1970s, which makes it a convention that was adopted because it worked.

IMRaD
Introduction, Methods, Results and Discussion.



Each section makes a promise, and each fails in a characteristic way.

The **title and abstract** promise the whole study in about 250 words. The abstract is worth reading first for orientation and again at the end, to check that it described what the paper actually contains. Abstracts are systematically more confident than the papers beneath them, and Section 6.5 gives a published example where the abstract quietly reports something other than what its own table shows.

The **Introduction** promises to explain why the question was worth asking, and should end by stating the objective in a sentence. When that sentence cannot be found, the fault usually lies with the authors, and a vague introduction tends to precede a vague analysis.

The **Methods** promise what was done, in enough detail to repeat it. This is the section that decides whether the Results are allowed to mean anything, and it is the one beginners skip.

The **Results** promise what happened, with no interpretation. What the Methods said would be measured is worth comparing against what actually appears here. Analyses that show up in the Results without having been promised in the Methods deserve attention.

The **Discussion** promises what the authors think it all means. That is opinion, and useful, and not evidence.

2.1 The order in which to read

Abstract, then Methods, then Results, then Introduction, then Discussion.

That ordering is deliberate. Results are seductive: a striking number persuades long before there is any way of knowing whether it was produced sensibly. Reading the Methods first means meeting every number already knowing how it was made, which is the difference between being informed and being convinced.

A second habit is harder to acquire. Many readers work through the abstract and the Discussion and feel they have read the paper. What they have read is what the authors would like them to conclude, which is a different object.

The point

Methods before Results. The Discussion is interpretation, not evidence. Those two habits do more for a reader in the first month than anything else in this chapter.

2.2 Where the numbers live

Table 1 reports who was in the study, broken down by group: age, sex, baseline measurements, medication. It answers one question, which is whether the groups were comparable before anything was done to them. In a randomised trial a large imbalance suggests either a small study or a problem with the randomisation. In an observational study imbalance is expected, and dealing with it is what the analysis will spend its time on.

Table 2 onwards reports what happened: the effect estimates, their confidence intervals, the *p* values. This is the material the abstract was summarising, and it is where a summary and its source can be checked against each other.

The flow diagram reports who entered the study, who was excluded and why, and who was lost along the way. It is often the most revealing figure in a paper and the easiest to skim past. A trial that loses a quarter of its participants, or loses them unevenly between the arms, has a problem that no amount of analysis will repair.

3 Framing an answerable question

A question written in four parts is already halfway to a study design. That format is called **PICO**, and it was set out in 1995 for clinicians framing questions at the point of care.

PICO
Population, Intervention, Comparison, Outcome.

Definition 3.1 • PICO

A clinical question is answerable when four components are specified. **P**, the population: who the question is about. **I**, the intervention or exposure: what is done to them, or what they are exposed to. **C**, the comparison: the alternative against which the intervention is judged. **O**, the outcome: what will be measured, in what units, at what time.

Each component corresponds to something the eventual analysis will need. P fixes the population that any conclusion may refer to. I and C together define the variable whose effect is being estimated, and crucially they define its two levels, because an effect is always an ef-

fect relative to something. O fixes the outcome variable, its unit and its time point, without which nothing can be counted.

Example 3.1 • COBRA-BPS decomposed

COBRA-BPS was a cluster-randomised trial of a community programme for high blood pressure, conducted in rural Bangladesh, Pakistan and Sri Lanka and published in 2020.

P: adults with hypertension living in thirty rural communities in Bangladesh, Pakistan and Sri Lanka.
I: a multicomponent programme, consisting of home visits by trained government community health workers for blood-pressure monitoring and counselling, training for physicians, and coordination of care in the public sector.
C: usual care.
O: reduction in systolic blood pressure, in mm Hg, at 24 months.

Example 3.1 has already done a great deal of work. The P identifies who the findings might apply to and, by omission, who they will not: the trial says nothing about urban patients or about anyone outside those three countries. The C shows that the comparison is against usual care rather than against nothing, so the result measures what the programme adds on top of what people already receive. The O carries a unit and a time point, which is what makes it measurable at all.

3.1 Where questions fail

A weak question fails in a predictable place, and it is almost always the same two letters.

The two components that get skipped

C and **O**.

Without a statement of what the comparison is, there is no study, only a description. “Blood pressure in this clinic” is a fact, not a finding.

Without a statement of exactly what will be measured, with a unit and a time point, there is no outcome, only a hope. “Improvement” is not an outcome. “Fall in systolic blood pressure in mm Hg at six months” is.

A first attempt at a PICO question is nearly always too broad. That is not a sign of inexperience so much as a sign that the question has not yet met the practical constraints that will narrow it: who can actually be recruited, what can actually be measured, and over what period anyone can afford to follow them.

4 Populations, samples, parameters and statistics

This section establishes the vocabulary that the rest of the chapter uses. It is short, and everything after it depends on it.

A study measures some people so as to say something about more people than it measured. Four objects are involved, and confusing any two of them produces most of the errors that appear in medical writing.

Definition 4.1 • Population, sample, parameter, statistic

The **population** is everyone the study is about. It is a decision made by the researcher, not a country or a census.

The **sample** is the part of the population actually measured.

A **parameter** is a numerical property of the population, such as the true mean μ or the true proportion π. It is almost always unknown.

A **statistic** is the corresponding property of the sample, such as $\bar{x}$ or $\hat{p}$. It is always known, because it was calculated from the data in hand.

Hat notation carries this distinction: π is the truth and $\hat{p}$ is the estimate of it that this particular study happened to produce. A different sample from the same population would produce a different $\hat{p}$. How much it would differ, and what follows from that, is the subject of Session 7.

Population and sample
The *population* is everyone the study is about, which is a decision the researcher makes. The *sample* is the part of it actually measured.

4.1 Statistical inference

Definition 4.2 • Statistical inference

The procedure by which a conclusion is reached about a *population* on the basis of information contained in a *sample* drawn from that population.

Three words in Definition 4.2 carry the weight. *Population*, because the conclusion is about people who were never measured. *Sample*, because the evidence is only ever partial. And *procedure*, because this is a method with rules rather than an impression, and the rules are what allow somebody else to check the work.

4.2 The test that separates the two

Take any number in a paper and ask one question of it: is this a fact about the people in this study, or a claim about people outside it?

Description	Inference
A fact about the sample	A claim about the population
Cannot be wrong, only miscalculated	Can be wrong even when calculated perfectly
Needs no assumptions	Needs assumptions about how the sample was drawn
2,645 adults; mean 146.7 mm Hg; 53.2% controlled	95% CI 3.2 to 7.1; RR 1.22; $P < 0.001$

Work through it with COBRA-BPS. The figure of 53.2 per cent is a fact. Somebody counted how many of those particular participants reached controlled blood pressure. The measurement procedure can be disputed; the count cannot.

By contrast, the interval from 3.2 to 7.1 mm Hg is not a fact at all. Nobody counted it. It is a statement about a large population of adults in rural South Asia whom nobody in that trial ever met, arrived at by assuming that the 2,645 people who were measured resemble them in the relevant respects. It is a bet, made carefully, with its odds written down.

Almost every serious argument about a paper is an argument about the second column rather than the first.

The commonest error in medical writing

A descriptive statement wearing the clothes of an inferential one.

"Mortality was lower in the treated group" is description. "The treatment reduces mortality" is inference, and it requires a great deal more than a lower number to support it.

Tense is often the only signal. Descriptions sit in the past, because they report what happened. Inferences slide into the present, because they claim something general. That shift of tense is frequently the only place where a much larger claim announces itself.

5 Reading a result: magnitude, precision, evidence

This is the most immediately useful skill in the chapter, and it repays practice until it is automatic. Here is the main result of COBRA-BPS, exactly as the paper reports it:

> *"the mean reduction was 5.2 mm Hg greater with the intervention (95% confidence interval [CI], 3.2 to 7.1; P < 0.001)"*

It reads as one claim. It is three, and they answer different questions.

	The number	The question it answers
Magnitude	5.2 mm Hg	How big is the effect?
Precision	95% CI 3.2 to 7.1	How well does this study pin that size down?
Evidence	$P < 0.001$	How badly do these data fit "no difference at all"?

5.1 Magnitude

That 5.2 is the estimate of how large the effect is, and it is the only one of the three expressed in clinical units. That makes it the only one open to direct clinical judgement, using what is already known about blood pressure and what it does to people over time.

It is also the number most often skipped. A reader trained to hunt for significance goes straight past the size of the effect to the *p* value, which is exactly backwards. Given a large enough study, an effect far too small to matter to anybody will still be statistically significant.

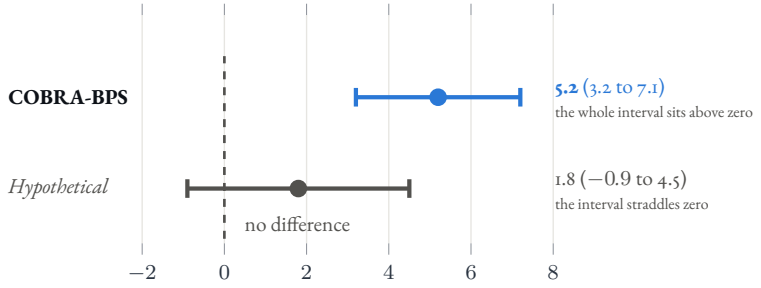
5.2 Precision

A result is a range, and the single headline number is only the middle of it.

Confidence interval
A range of values for the true effect that are reasonably consistent with what this study observed.

An interval indicates how far the estimate could be from the truth. Read the COBRA-BPS interval as follows: given what this trial observed, the true difference is plausibly anywhere between 3.2 and 7.1 mm Hg. A narrow interval indicates a precise study, usually a large one. A wide interval indicates that the study could not pin the answer down, whatever its headline number says. The exact definition of what 95 per cent refers to, and how the interval is constructed, is Session 8; the operational reading above is enough for now and is not misleading.

What changes a reader's habits is looking at *where* the interval sits, not only at how wide it is.



Both ends of the COBRA-BPS interval, 3.2 and 7.1, lie above zero, and zero is the value that would mean no difference. Every value the data find plausible is a benefit, and that is what licenses the authors' claim of an effect.

Below it sits a hypothetical study reporting a difference of 1.8 mm Hg with an interval running from −0.9 to 4.5. The best estimate is still a benefit. But the interval now includes zero, and it includes small harms alongside moderate benefits. That study has not shown that the treatment fails. It has shown that it could not tell.

The most useful distinction in this chapter

"No evidence of effect" and "evidence of no effect" are different statements, and papers blur them constantly.

When a study reports no significant difference, the conclusion that the treatment fails does not follow. Two questions settle what has actually been learned: how large was the study, and how wide is the interval? An interval that contains both nothing and a

large benefit has reported nothing at all.

5.3 Evidence, and what the *p* value is for

Imagine a world in which the intervention does absolutely nothing. In that world, could a 5.2 mm Hg difference still appear between the two groups, purely through the luck of who ended up in which one? Yes, sometimes. Small imbalances happen by chance.

A *p* value answers exactly one question: how often. Here, less than once in a thousand times. So either something real is going on, or something quite unusual happened in this particular trial.

That is the whole of what it says, and it is much less than most readers assume.

What the *p* value does not tell you

It is not the probability that the treatment works. It is not the probability that the result is a fluke. It says nothing about how large the effect is, and nothing about whether an effect that size is worth having.

There is also nothing special about 0.05. Fisher offered it as a rough working threshold and said as much himself. Treating 0.049 and 0.051 as categorically different is indefensible, and in 2016 the American Statistical Association published a formal statement to that effect, because the misuse had become serious enough to warrant one.

5.4 The order to read them in

Magnitude first, interval second, *p* value last. The first says whether the effect would matter clinically, the second says how firmly the study established it, and the third says only whether chance alone is a comfortable explanation. Reversing that order is how a trivial effect in a very large study comes to be reported as an important finding.

6 Effect measures: absolute and relative

The result in Section 5 was a difference between two means. When the outcome is instead something a patient either does or does not experience, reaching controlled blood pressure, attending a referral appointment, dying within a year, the comparison is between two proportions. There is more than one way to express the difference between two proportions, the ways give very different impressions of the same data, and the choice is not neutral.

6.1 The two-by-two table and its notation

Everything in this section comes out of one small table.

	Outcome occurs	Outcome does not	Total
Treated group	a	b	$n_1 = a + b$
Comparison group	c	d	$n_0 = c + d$

In each group, the proportion experiencing the outcome, usually called the risk, is the count divided by the group total:

$$\hat{p}_1 = \frac{a}{n_1}, \qquad \hat{p}_0 = \frac{c}{n_0}. \tag{6.1}$$

a, c the number with the outcome in the treated and comparison groups
n_1, n_0 the total number of people in each group
$\hat{p}_1$, $\hat{p}_0$ the observed proportion, or risk, in each group

One trial runs through the rest of this section. It was a randomised trial conducted in Barishal, Bangladesh, and published open access in 2020. It tested whether a health education package increased the proportion of patients with probable diabetic retinopathy who acted on a referral. Its primary outcome table gives the four counts directly.

Example 6.1 • Referral compliance, from the trial's own table

	Complied	Did not comply	Total
Health education	92	51	143
Standard care	44	112	156

Applying (6.1) gives $\hat{p}_1 = 92/143 = 0.643$ and $\hat{p}_0 = 44/156 = 0.282$, which the paper reports as 64.3 per cent and 28.2 per cent.

6.2 The absolute measures

The **risk difference** subtracts one proportion from the other:

$$\widehat{\mathrm{RD}} = \hat{p}_1 - \hat{p}_0. \tag{6.2}$$

It is expressed in percentage points, and it answers the question a health ministry asks: out of every hundred people put through this programme, how many extra reach the outcome?

For the referral trial, $\widehat{\mathrm{RD}} = 0.643 - 0.282 = 0.361$, that is 36.1 percentage points, which is also the figure the paper itself reports. For COBRA-BPS, where 53.2 per cent of the intervention group reached controlled blood pressure against 43.7 per cent of the control group, the same subtraction gives 9.5 percentage points.

The **number needed to treat** is the reciprocal of the risk difference:

$$\widehat{\mathrm{NNT}} = \frac{1}{|\widehat{\mathrm{RD}}|}. \tag{6.3}$$

It converts the risk difference into a count of people, which is often the most concrete form a result can take.

Intuition

If treating a hundred people produces nine extra good outcomes, then producing one extra good outcome takes about a hundred divided by nine, which is roughly eleven people. Equation (6.3) is that arithmetic written once and for all. A small risk difference therefore produces a large NNT, and an NNT of 200 is a way of saying that almost everyone treated receives no benefit.

For COBRA-BPS, $1/0.095 = 10.5$, so roughly eleven people go through the programme for one additional person to reach controlled blood pressure. For the referral trial, $1/0.361 = 2.8$, so roughly three people receive the education for one additional person to attend. An NNT near three is unusually small, which is worth a moment of suspicion rather than celebration; Section 7 and the exercises return to why.

6.3 The relative measures

The **relative risk**, also called the risk ratio, divides instead of subtracting:

$$\widehat{\mathrm{RR}} = \frac{\hat{p}_1}{\hat{p}_0}. \tag{6.4}$$

A relative risk of 1 means no difference. Above 1 the outcome is more common in the treated group, below 1 it is less common.

COBRA-BPS reports a relative risk of 1.22, with a 95 per cent confidence interval of 1.10 to 1.35. Dividing the two published percentages, $53.2/43.7 = 1.22$, reproduces it, which is a useful check that the quantity has been correctly understood. Applying (6.4) to the referral trial gives $0.643/0.282 = 2.28$: patients who received the education were about 2.3 times as likely to attend.

One result therefore has two honest expressions. The COBRA-BPS programme raised control rates by 9.5 percentage points, and it raised them by 22 per cent relative to usual care. Both statements describe the identical pair of numbers.

Watch out

Identical data, two framings. The relative figure sells a programme; the absolute figure lets a ministry cost it. Neither is dishonest on its own, but presenting only the relative number usually is, because a large relative change on a small baseline risk is compatible with an almost negligible absolute benefit. Wherever a relative figure appears alone, the absolute one is the thing to ask for.

6.4 Odds, and the odds ratio

The **odds** of an outcome is the probability that it happens divided by the probability that it does not:

$$\widehat{\mathrm{odds}} = \frac{\hat{p}}{1-\hat{p}} = \frac{\text{number with the outcome}}{\text{number without it}}. \tag{6.5}$$

The **odds ratio** compares the odds in the two groups:

$$\widehat{\mathrm{OR}} = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_0/(1-\hat{p}_0)} = \frac{a\,d}{b\,c}. \tag{6.6}$$

a, *b* with and without the outcome, treated group
c, *d* with and without the outcome, comparison group

Odds ratios are everywhere in the medical literature, for reasons that belong to Session 25: it is what logistic regression produces, and it is the only valid effect measure in a case-control study. It is also routinely misread as though it were a relative risk, and the two are not the same number.

For the referral trial, (6.6) gives

$$\widehat{\text{OR}} = \frac{92 \times 112}{51 \times 44} = \frac{10{,}304}{2{,}244} = 4.59,$$

against a relative risk of 2.28 from (6.4). The odds ratio is twice the size of the risk ratio, and both describe the same four counts.

Intuition

Odds ratio and risk ratio agree closely when the outcome is rare, because when $\hat{p}$ is small, $1 - \hat{p}$ is close to 1 and dividing by it changes little. As the outcome becomes common the denominators start to matter and the odds ratio moves away from the risk ratio, always in the direction of looking larger. Here the outcome occurred in 28 per cent of the control group, which is not rare at all, so the gap is wide. An odds ratio of 4.59 quoted as though it meant "4.6 times as likely to attend" would overstate the finding by a factor of two.

6.5 Crude and adjusted, and why the distinction is not academic

The referral trial's abstract reports its main result as "64.3% vs 28.2%; OR 4.73; 95% CI 2.87–7.79; $p < 0.001$".

The odds ratio computed above from the trial's own primary outcome table is 4.59, not 4.73. That difference is not a rounding error. Reading the Methods and Table 6 shows that 4.73 is an **adjusted** odds ratio, produced by a multivariable logistic regression that also included the participant's self-perception of a vision problem and their monthly income. The 4.59 is the **crude** odds ratio, which compares the two groups and nothing else.

Neither number is wrong. They are answers to different questions. The crude figure asks what happened in the two groups as randomised. The adjusted figure asks what the education contributed once two other predictors of attendance were also in the model. What the abstract does is place the adjusted figure immediately after the two crude percentages, where a quick reader will take it for the comparison of those two. This is a good example of why Section 2 recommends reading an abstract twice, and checking it against the tables in between.

A habit worth forming immediately

An odds ratio in an abstract carries no label saying whether it is crude or adjusted, and if it is adjusted, for what. Those facts live in the Methods and in the regression table. A reader who does not go and look cannot know which of the two quantities is being quoted.

The same trial illustrates a smaller point in the same direction. The lower limit of the confidence interval appears as 2.87 in the abstract and as 2.89 in Table 6. Nothing of consequence turns on it, and it is a reminder that the tables, not the abstract, are the record.

6.6 Keeping provenance visible

Every figure used in this section appears below, with a statement of where each one came from. That right-hand column is the point of it.

Measure	Value	Where it comes from
COBRA-BPS, blood-pressure control		
Control rate, intervention	53.2%	reported in the paper
Control rate, usual care	43.7%	reported in the paper
Relative risk	1.22 (1.10 to 1.35)	reported in the paper
Risk difference	9.5 points	our subtraction
Number needed to treat	about 11	our division, 1/0.095
Referral trial, compliance		
Compliance, education group	64.3% (92/143)	reported in the paper
Compliance, standard care	28.2% (44/156)	reported in the paper
Risk difference	36.1 points	reported in the paper
Relative risk	2.28	our division
Crude odds ratio	4.59	our calculation from Table 2
Adjusted odds ratio	4.73 (2.87 to 7.79)	reported, from logistic regression
Number needed to treat	about 3	our division, 1/0.361

Six of these thirteen figures were calculated here rather than printed in either paper. Marking which is which is not pedantry. It is the difference between reporting a study and reworking it, and a reader who cannot tell the two apart in someone else's writing will not manage it in their own. Journals expect the habit in anything submitted for publication.

On the odds ratio

An odds ratio is not a relative risk. It exceeds the relative risk whenever the outcome is common, and in an abstract it is often adjusted for variables that abstract never names. All three facts are ordinary, and all three are routinely missed.

7 What statistics cannot repair

It would be reasonable to finish a first course thinking that statistics is what happens to data after collection. That is the most expensive misunderstanding in research, and it is worth correcting before anyone designs a study of their own.

7.1 A bad question

No analysis rescues a question that was vague, unanswerable, or of no interest to anyone. Beginners tend toward vague questions because they feel safer, and a claim that says nothing cannot be refuted. It also cannot be answered. Section 3 is the remedy, and it works in proportion to how uncomfortably specific the four components are made.

7.2 A biased sample

Bias is systematic error, and systematic error does not average out. This is the part that surprises people: a *larger* sample does not help. If the people studied differ systematically from the people the conclusion is about, collecting more of them narrows the confidence interval around the wrong answer. What results is greater confidence in something untrue, which is worse than uncertainty.

That unusually small NNT in the referral trial is worth pausing on here. The participants were people already registered at a diabetes hospital who had been referred and had not attended. That is a group selected for being reachable, and an education programme delivered to them says less about people who never reached a diabetes hospital at all. Its authors state this in their own limitations.

7.3 A bad measurement

A poor measurement makes everything downstream meaningless, however immaculate the analysis. Two properties are confused here often enough to be worth separating.

Definition 7.1 • Reliability and validity

Reliability is consistency: the same instrument, applied twice under the same conditions, returns the same answer.

Validity is correctness: the instrument measures the quantity it is claimed to measure.

A bathroom scale reading three kilograms heavy every single time is perfectly reliable and completely invalid. These two properties are independent, and reliability is the easier to demonstrate, which is why it is more often the one reported. Session 14 takes this apart properly.

The point

Almost everything determining whether a study is worth anything is settled *before the first patient is enrolled*: the question, the design, the population, the outcome, the sample size. By the time a spreadsheet exists, the decisions that mattered have been made.

One professional habit follows immediately. Consult a statistician before starting, not afterwards. A statistician most often cannot help because the person arrived too late for the answer to change anything.

8 A framework for appraisal

Five questions, applicable to any clinical paper ever written, and answerable in about twenty minutes.

1. **What was the question?**
   If it cannot be stated in one sentence, either it has not been found or the authors never had one.
2. **Who was studied, and who was left out?**
   Generalisability is decided in the Methods, not in the Discussion.
3. **What was compared with what?**
   No comparison, no evidence. Find the control group; where there is none, say so.
4. **How big was the effect, and how precise?**
   The estimate *and* the interval, in absolute as well as relative terms. Never the p value alone.
5. **What could explain this other than the intervention?**
   Bias, confounding, chance. This is the question that separates a reader from an appraiser.

9 Chapter summary

1. The effect of a treatment on an individual patient, $\tau_i = Y_i(1) - Y_i(0)$, can never be observed, because only one of the two outcomes ever happens. Research estimates the average effect across a group instead, and every study design is an attempt to make the comparison group a believable stand-in.
2. Authority and personal experience both fail as grounds for believing a treatment works: the first because the chain may end nowhere, the second because the counterfactual is missing from every case seen.
3. Papers are built as IMRaD. Read the Methods before the Results, and treat the Discussion as interpretation.
4. A question becomes answerable when its population, intervention, comparison and outcome are all specified. The comparison and the outcome are the two that get skipped.
5. Description is a fact about the sample and cannot be wrong. Inference is a claim about the population and can be wrong even when the arithmetic is perfect.
6. Every result carries three separate claims: magnitude, precision, evidence. Read them in that order.
7. An interval that crosses the null value means the study could not tell, which is not the same as showing that the treatment does not work.

8. The same pair of proportions can be reported as a risk difference, a relative risk, an odds ratio or a number needed to treat. Relative forms look larger. Absolute forms are the ones a health system can act on.
9. An odds ratio exceeds the relative risk whenever the outcome is common, and an odds ratio quoted in an abstract may be adjusted for variables the abstract never names.
10. The question, the design, the sampling and the measurement are all settled before data exist, and no analysis repairs them afterwards.

10 Key terms

Absolute risk difference
The difference between two proportions, in percentage points. See (6.2).

Allocation concealment
Preventing anyone from knowing which arm the next patient will enter. Not the same as blinding.

Average treatment effect
The mean effect across a population, $\mathrm{E}[Y(1)] - \mathrm{E}[Y(0)]$. The quantity studies actually estimate.

Bias
Systematic error. Does not shrink with a larger sample.

Confidence interval
A range of values for the true effect consistent with the data.

Confounding
A third factor linked to both exposure and outcome, creating an association that is not causal.

Control group
The comparison group. The stand-in for the counterfactual.

Counterfactual
What would have happened to the same patient under the other treatment. Never observed.

Crude estimate
An effect estimated from the comparison alone, with no adjustment for other variables.

Adjusted estimate
An effect estimated from a model that also contains other variables. Not interchangeable with the crude one.

Descriptive statistics
Summaries of the data actually collected.

Effect size
How large the difference or association is, in interpretable units.

IMRaD
Introduction, Methods, Results and Discussion.

Inferential statistics
Conclusions about a population, drawn from a sample.

NNT
Number needed to treat. The reciprocal of the absolute risk difference. See (6.3).

Null hypothesis
The proposition that there is no effect.

Odds
The probability an outcome happens divided by the probability it does not. See (6.5).

Odds ratio
The ratio of the odds in two groups. See (6.6).

Parameter
A property of a population, such as μ or π. Usually unknown.

PICO
Population, Intervention, Comparison, Outcome.

Population
Everyone the study is about.

***p* value**
The probability of data at least as extreme as those observed, if the null hypothesis were true.

Randomisation
Allocation by chance, so that groups are comparable in respects both measured and unmeasured.

Relative risk
The ratio of risk in one group to risk in another. See (6.4).

Reliability
Consistency of measurement.

Sample
A part of a population.

Statistic
A property of a sample, such as $\bar{x}$ or $\hat{p}$. Always known.

Validity
Whether an instrument measures what it claims to.

11 Exercises

Twenty exercises, taking roughly an hour. Worked solutions are in Section 12, and they explain the reasoning rather than just stating the verdict. Attempting each one first is the point.

Parts C to F use one real paper, printed below. It is a randomised trial conducted in Barishal, Bangladesh, and published open access. Read it once before starting Part C. The counts in Example 6.1 come from Table 2 of the same paper.

THE PAPER

Khair Z, Rahman MM, Kazawa K, Jahan Y, Faruque ASG, Chisti MJ, Moriyama M. Health education improves referral compliance of persons with probable diabetic retinopathy: a randomized controlled trial. *PLoS ONE* 2020;15(11):e0242047.

Objective. Lack of awareness about diabetic retinopathy (DR) is the most commonly cited reason why many persons with type 2 diabetes are non-compliant with referral instruction to undergo retinal screening. The purpose of this study was to evaluate the efficacy of a culturally, geographically and socially appropriate, locally adapted five-month-long health education on referral compliance of participants.

Method. A prospective randomized, open-label parallel group study was conducted on persons with type 2 diabetes who underwent basic eye screening at a diabetes hospital between September 2017 and August 2018. Participants who were noncompliant with referral instruction to visit a hospital for advanced DR management were randomly divided into health education intervention group ($n = 143$) and control group ($n = 156$). Both groups received information regarding DR and referral instruction at the diabetes hospital. The intervention group was provided personalized education followed by telephonic reminders. The primary endpoint was 'increase in referral compliance' and the secondary endpoint was 'increase in knowledge of DR'. Multivariate logistic regression model was used to identify significant predictors of compliance to referral.

Results. A total of nine participants dropped and 290 completed the post intervention survey. The compliance rate in intervention group was found to be significantly higher than the control group (64.3% vs 28.2%; OR 4.73; 95% CI 2.87–7.79; $p < 0.001$).

 Study site: Barishal district, Bangladesh. Registered as NCT03658980 and approved by the Bangladesh Medical Research Council.

11.1 Part A. Description or inference

For each statement, decide whether it is a **description**, a fact about the people studied, or an **inference**, a claim about people who were not studied. Give one line of reasoning.

1. "143 participants were allocated to the intervention group."

2. "The compliance rate was 64.3% in the intervention group."

3. "The odds ratio was 4.73, with a 95% confidence interval of 2.87 to 7.79."

4. "Health education improves referral compliance in people with diabetic retinopathy."

5. "Mean age was 51.5 years in the intervention group and 51.0 years in the control group."

11.2 Part B. Reading a result sentence

6. In "OR 4.73; 95% CI 2.87–7.79; $p < 0.001$", identify which number is the magnitude, which the precision, and which the evidence.

7. Suppose a different study of the same intervention reported "OR 1.40; 95% CI 0.85–2.30; $p = 0.19$". Does it follow that health education does not work? Answer in two sentences.

8. For a ratio measure such as an odds ratio or a relative risk, the value meaning "no difference" is not zero. What is it, and why?

11.3 Part C. Extract the PICO

9. Fill in all four components from the abstract above. Be specific: a vague answer here is the commonest mistake.

P Population

I Intervention

C Comparison

O Outcome

11.4 Part D. Calculation

Use the four counts in Example 6.1: 92 of 143 complied in the education group, 44 of 156 in the standard care group. Show the arithmetic in each case.

10. Compute $\hat{p}_1$ and $\hat{p}_0$ using (6.1), and confirm that they match the percentages the paper reports.

11. Compute the absolute risk difference using (6.2), and the number needed to treat using (6.3). State what the NNT means in a sentence a hospital administrator would understand.

12. Compute the relative risk using (6.4). Express the same finding twice, once in absolute and once in relative terms.

13. Compute the crude odds ratio using (6.6). It is roughly twice the relative risk. Explain why, referring to the size of $\hat{p}_0$.

14. COBRA-BPS reported all-cause mortality of 2.9 per cent in the intervention group and 4.3 per cent in the control group. Compute the risk difference, the relative risk and the odds ratio for this outcome. Compare the relative risk with the odds ratio, and explain why the relationship between them differs from the one found in the previous exercise. *(Note before concluding anything: mortality was not the primary outcome and the paper reports no confidence interval for it.)*

15. The crude odds ratio calculated in exercise 13 is not the 4.73 the abstract reports. Where does the difference come from, and which of the two numbers answers the question "what happened in the two groups as randomised"?

11.5 Part E. Does the conclusion stay inside the data?

Mark each statement **fair** or **overreach**, and say why.

16. "Personalised health education increased referral compliance in this population."

17. "Health education prevents blindness from diabetic retinopathy."

18. "These findings apply to all people with diabetes in Bangladesh."

11.6 Part F. Appraisal

19. Apply questions 1 to 3 from Section 8 to this paper. One sentence each.

20. Question 5 is the hard one. Name **two** things other than the health education itself that could have produced a higher compliance rate in the intervention group. *(The trial was open-label, and the intervention group also received telephone calls.)*

12 Solutions

Part A

1. Description. A count of what was done in this study.

2. Description. Somebody counted how many of those 143 people attended. It is a fact about them.

3. Inference. The odds ratio summarises this sample, but the confidence interval attached to it is a statement about the wider population of similar patients who were never in the study.

4. Inference, and the tense is the giveaway. "Improves", in the present, is a general claim about people beyond this trial. Compare "improved compliance in this study", which would be description.

5. Description. This is Table 1 material: who was in the study. It is there so that a reader can judge whether the two groups started out comparable.

Part B

6. Magnitude is **OR 4.73**, precision is **95% CI 2.87 to 7.79**, evidence is $p < 0.001$.

7. No. It follows only that *this study could not tell*. The interval runs from 0.85 to 2.30, which includes a modest harm and a substantial benefit alike, so the data are compatible with too many different truths to single one out. Absence of evidence is not evidence of absence.

8. The null value for a ratio is **1**, not 0, because a ratio of two equal quantities is 1. So a confidence interval for an odds ratio or a relative risk indicates no detectable effect when it contains 1, in the same way that an interval for a difference does when it contains 0. Here, 2.87 to 7.79 lies entirely above 1.

Part C

9. P: adults with type 2 diabetes, registered at a diabetes hospital, who had been referred for advanced DR management and had not complied with that referral.
I: a five-month locally adapted health education package, consisting of a personalised face-to-face session followed by telephone reminders.
C: usual care, being the standard DR information and referral instruction given at the diabetes hospital, with no personalised education.
O: referral compliance, that is, whether the participant attended the tertiary hospital. (Secondary: knowledge of DR.)

A blank C, or one reading "nothing", is worth looking at again. That control group did receive something. The trial compares education *added to* usual care, not education against

nothing at all.

Part D

10. $\hat{p}_1 = 92/143 = 0.643$ and $\hat{p}_0 = 44/156 = 0.282$, which are the 64.3 per cent and 28.2 per cent the paper reports. Reproducing a paper's own percentages from its counts is a cheap and worthwhile check that the denominators are what they appear to be.

11. $\widehat{\mathrm{RD}} = 0.643 - 0.282 = 0.361$, that is **36.1 percentage points**, which the paper also reports. $\widehat{\mathrm{NNT}} = 1/0.361 = 2.77$, so about **three**. Stated usefully: for every three patients put through the education programme, roughly one additional patient attends the referral appointment who would not otherwise have done so.

12. $\widehat{\mathrm{RR}} = 0.643/0.282 = \mathbf{2.28}$. In absolute terms, the programme raised attendance by 36.1 percentage points. In relative terms, it made attendance about 2.3 times as likely. Both sentences describe the same four counts.

13. $\widehat{\mathrm{OR}} = (92 \times 112)/(51 \times 44) = 10{,}304/2{,}244 = \mathbf{4.59}$, against a relative risk of 2.28.

That comes down to the size of $\hat{p}_0$. The odds ratio divides each proportion by one minus itself before comparing them. When a proportion is small that division changes almost nothing, so the odds ratio and the relative risk nearly coincide. Here the control group proportion is 0.282, which is not small, and the intervention group proportion of 0.643 is larger still, so both divisions matter and they pull in opposite directions. The odds ratio is inflated relative to the risk ratio, which is what always happens when the outcome is common.

14. With $\hat{p}_1 = 0.029$ and $\hat{p}_0 = 0.043$: risk difference $= -0.014$, that is **1.4 percentage points fewer** deaths in the intervention group; $\widehat{\mathrm{RR}} = 0.029/0.043 = \mathbf{0.67}$; $\widehat{\mathrm{OR}} = (0.029/0.971) \div (0.043/0.957) = \mathbf{0.66}$.

Here the relative risk and the odds ratio agree to within 0.01, whereas in the previous exercise they differed by a factor of two. The difference is that death occurred in about 3 to 4 per cent of participants, so $1 - \hat{p}$ is close to 1 in both groups and dividing by it barely moves anything. This is the rare-outcome case in which the odds ratio approximates the relative risk.

Two cautions. The NNT here would be $1/0.014 \approx 71$. None of these figures is evidence that the programme reduced mortality. Mortality was not the primary outcome and carries no confidence interval, and a trial sized to detect a blood-pressure difference is not sized to detect a difference in deaths.

15. The 4.59 is the **crude** odds ratio, computed from the two groups alone. The 4.73 is an **adjusted** odds ratio from a multivariable logistic regression that also included self-perception of a vision problem and monthly income. The question "what happened in the two groups as randomised" is answered by the **crude** figure, 4.59.

This matters beyond arithmetic. The abstract places 4.73 directly after the two crude percentages, where it reads as a comparison of them, and nothing in the abstract says the figure is adjusted or names what it was adjusted for. Only the Methods and Table 6 make that clear.

Part E

16. Fair. It reports what happened and limits itself to this population.

17. Overreach. The outcome measured was attendance at a referral appointment, not vision. Attending may well lead to earlier treatment and less blindness, but this study did not measure that. Quietly upgrading the outcome that was measured to the outcome everybody cares about is one of the commonest ways a paper overreaches.

18. Overreach. The participants were already registered with a diabetes hospital, so they were more aware of their condition than the general diabetic population, and the referral hospital was in the same district. The authors state both points themselves in their limitations.

Part F

19. 1. The question: does a locally adapted health education package increase the proportion of referred patients with type 2 diabetes who actually attend for advanced DR management?
2. Who was studied: adults referred from one diabetes hospital in Barishal who had not complied. Left out: people not registered at that hospital, people never referred, and anyone under 18.
3. What was compared: personalised education plus telephone reminders, against the standard information both groups received anyway.

20. Several answers are reasonable. Two good ones:

The telephone calls rather than the education. The intervention was a package: a face-to-face session *and* reminder calls. A reminder alone might account for much of the effect. The trial cannot separate the components, which is a limitation of any multicomponent intervention.

Attention rather than content. Being taught in person and telephoned repeatedly may change behaviour regardless of what was said. The trial was open-label, so participants knew they were receiving something extra.

Also creditable: the intervention group had more contact time overall; and compliance was ascertained at the hospital, so anyone attending for an unrelated reason might be recorded as compliant.

None of this makes the trial a bad one. Every trial has such alternatives. Noticing them is the skill being practised, which is question 5 and the whole of Sessions 12 to 15.

13 References and further reading

Every item was identified through PubMed and checked against the record there. Where a paper is free to read, the PubMed Central identifier is given.

The two studies used throughout

1. Jafar TH, Gandhi M, de Silva HA, Jehan I, Naheed A, et al. A community-based intervention for managing hypertension in rural South Asia. *N Engl J Med* 2020;382(8):717–726. doi:10.1056/NEJMoa1911965. PMID 32074419. Paywalled; only the published abstract is quoted here, and its numbers are re-typeset rather than reproduced.
2. Khair Z, Rahman MM, Kazawa K, Jahan Y, Faruque ASG, Chisti MJ, Moriyama M. Health education improves referral compliance of persons with probable diabetic retinopathy: a randomized controlled trial. *PLoS ONE* 2020;15(11):e0242047. doi:10.1371/journal.pone.0242047. PMID 33180863. Free: PMC7660573. Published under CC BY 4.0, which is why its abstract can be reproduced in Section 11.

On method and reporting

3. Richardson WS, Wilson MC, Nishikawa J, Hayward RS. The well-built clinical question: a key to evidence-based decisions. *ACP J Club* 1995;123(3):A12–13. doi:10.7326/ACPJC-1995-123-3-A12. PMID 7582737. The origin of PICO.
4. Sollaci LB, Pereira MG. The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey. *J Med Libr Assoc* 2004;92(3):364–367. PMID 15243643. Free: PMC442179. The source for the claim that IMRaD became near-universal only in the 1970s.
5. Altman DG, Bland JM. Absence of evidence is not evidence of absence. *BMJ* 1995;311(7003):485. doi:10.1136/bmj.311.7003.485. PMID 7647644. Free: PMC2550545. One page, and the best statement of the distinction in Section 5.
6. Wasserstein RL, Lazar NA. The ASA statement on *p*-values: context, process, and purpose. *The American Statistician* 2016;70(2):129–133. doi:10.1080/00031305.2016.1154108. Not indexed in PubMed. The formal statement referred to in Section 5.
7. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *PLoS Med* 2010;7(3):e1000251. doi:10.1371/journal.pmed.1000251. PMID 20352064. Free: PMC2844794.

Historical

8. Doll R, Hill AB. Smoking and carcinoma of the lung; preliminary report. *Br Med J* 1950;2(4682):739–748. doi:10.1136/bmj.2.4682.739. PMID 14772469. Free: PMC2038856. The paper that created the case-control design.
9. Medical Research Council. Streptomycin treatment of pulmonary tuberculosis. *Br Med J* 1948;2(4582):769–782. Reproduced by the James Lind Library.
10. Tulchinsky TH. John Snow, cholera, the Broad Street pump; waterborne diseases then and now. In: *Case Studies in Public Health*. Elsevier, 2018. doi:10.1016/B978-0-12-804571-8.00017-2. Free: PMC7150208.
11. Snow J. *On the Mode of Communication of Cholera*, 2nd ed. London: John Churchill, 1855. Public domain.

The Semmelweis mortality figures quoted in Section 1, 98.4 and 36.2 deaths per thousand births, are his own tabulations as reproduced by the James Lind Library. The Doll and Hill paper is recommended here for its design rather than its numbers: the surviving scans do not permit its cell counts to be quoted with confidence, so they are not quoted.

14 For students in the mentorship programme

Everything above stands on its own. This section is the only part tied to a particular session of the Stat.BD programme, and it can be ignored by anyone reading the chapter on its own.

Before the next session

1. Write **one** clinical question from your own practice, in PICO form, one sentence per component. It will probably be too broad. Every first attempt is. Bring it anyway.
2. Search PubMed for **one** paper that answers part of it. Any paper will do; searching properly is Session 11.
3. Run the **five questions** of Section 8 over it and bring the answers.

Reading

Doll and Hill 1950, reference 8. above, free at PMC2038856. Read it for the **design** rather than the numbers, watching how the control group was built. It is short, and it is the paper that created the case-control study.

Optional, and one page: Altman and Bland 1995, reference 5. above.

Questions this chapter deliberately left open

Several ideas were used before they were properly defined, because using them first is the faster way in. Each is settled later.

Question	Where it is answered
What does 95 per cent in a confidence interval actually refer to, and how is the interval built?	Session 8
How much would the estimate change in a different sample?	Session 7
Why is 0.05 the conventional threshold?	Session 8
What is the difference between a confounder, a mediator and an effect modifier?	Session 15
Relative risk, odds ratio, risk difference and NNT, treated properly	Session 23
What a logistic regression is doing when it produces an adjusted odds ratio	Session 25
Why cluster randomisation changes the analysis	Sessions 13 and 26
Reliability and validity of a measurement	Session 14

Next

Session 2: Population, Samples, Variables and Data.