Foundations and disease occurrence

Epidemiology: Study of the **distribution and determinants** of health states/events in populations, and its application to control health problems.

Distribution / determinants: Who, where and when / factors associated with or influencing outcomes.

Population / sample: The units we want conclusions about / the units actually studied.

Exposure / outcome: The factor investigated / the disease, event or state of interest. Exposure may or may not cause the outcome. Their roles depend on the research question.

Case definition: Criteria for deciding who has the outcome. Define the **case, population and time frame** before measuring frequency.

Morbidity / mortality: Illness / deaths in a population.

Endemic / epidemic: Usual, continuing presence / occurrence above the expected level in a defined place and period.

Point / period prevalence: Proportion affected **at one moment** / **at any time during a period**. Period prevalence includes existing and new cases; count each person once.

Incidence risk / cumulative incidence (CI): New cases divided by initially disease-free people over a specified period: a probability between 0 and 1.

Incidence rate (IR): New cases divided by **person-time at risk**. Measures speed of occurrence, not a probability.

Population at risk / person-time: People able to develop the outcome / their combined observation time before onset, loss, death or study end.

Odds: Cases divided by non-cases, unlike risk, which divides by everyone initially at risk.

Prevalence depends on incidence and duration. Longer survival can raise prevalence without raising incidence. Cure reduces duration. Prevalence helps plan services; incidence helps investigate causes.

Population data

Demography / population pyramid: Population size, structure and change / a graph of its age and sex structure, often in five-year age groups. Shape does not determine total size.

Census / registry / ICD: Population information at a specified time / systematic records of defined cases or events / International Classification of Diseases.

Crude / total mortality: All deaths relative to population size, without age adjustment. Age-specific rates use only that age group; cause-specific rates count deaths from one cause.

Life expectancy at birth: Expected years lived if the current **age-specific mortality pattern stayed unchanged**, not the average age of people dying this year.

Direct age-standardisation: Apply age-specific rates to a **common standard age distribution**. Need age-specific rates (or events and population counts) and standard weights. Compare the same outcome, units and standard.

Study designs

Descriptive / analytic: Describes patterns / examines associations through comparisons.

Observational / interventional: Measures existing exposures / researchers deliberately assign an intervention. An intervention is not automatically randomised.

Ecological study: Exposure and outcome measured for **groups**, such as countries. Often uses routinely collected, aggregated data.

Ecological fallacy: Assuming a group-level association necessarily applies to individuals. A country correlation does not link each person's exposure to their outcome.

Cross-sectional: Measures exposure and/or outcome at a specified time. Gives prevalence and current burden; usually cannot establish which came first.

Cohort: Classify by exposure, then follow outcomes. For first-event incidence, people start free of **that outcome**, not necessarily every illness.

Prospective / retrospective cohort: Observe subsequent outcomes after recruitment / reconstruct exposure and later outcomes from existing records. Both follow exposure-to-outcome logic.

Case-control: Select by **outcome status**, then compare previous exposures. Cases have the condition; controls represent the source population producing those cases.

Controls: Need not be completely healthy. They must represent the population producing the cases, not just be easy to recruit.

Design	Particularly useful for	Main limitation
Cohort	Rare exposures; several outcomes	Rare disease needs large samples or long follow-up.
Case-control	Rare disease; several past exposures	Recall and selection bias; usual sampled totals do not give population incidence.
Cross-sectional	Current burden and service needs	Uncertain time order; prevalence depends on duration.
Ecological	Group patterns and hypotheses	Individual conclusions may be invalid.

Asking about the past does not define case-control. A survey can ask about past exposure; a cohort can use old records. Identify how participants were selected.

Association measures

RR / rate ratio / prevalence ratio (PR, PRR): Compare risks / incidence rates / existing disease prevalence. These denominators are not interchangeable.

Odds ratio (OR): Ratio of odds. In case-control studies, compare exposure odds in cases and controls, not sampled disease risks.

Relative / absolute difference: Divide measures / **subtract** measures. No-difference value: **1** for ratios, **0** for differences.

Excess risk / ARR: Exposure comparisons may use exposed minus unexposed. Treatment benefit uses control adverse risk minus treated adverse risk. Always state the subtraction order.

Bias and causal interpretation

Association / causality: A statistical relationship / a cause-and-effect relationship. Association can instead result from chance, bias or confounding.

Temporality / reverse causation: A cause must precede its effect. Reverse causation means disease changes the apparent exposure, for example illness causing inactivity.

Bias / random error: Systematic distortion / chance variation. Larger samples improve precision but do not automatically remove bias.

Selection bias: Systematic differences in who enters or stays can distort results. Healthy volunteers may misrepresent disease frequency.

Loss to follow-up / attrition bias: Loss related to exposure and outcome can bias comparisons. No further observation does not mean the person stayed disease-free.

Information bias: Systematic error measuring or recording exposure/outcome. Different accuracy between groups can create or hide associations.

Recall / observer bias: Groups remember past exposure differently / expectations influence questioning, measurement or recording. Both are information bias.

Misclassification: Wrong exposure or outcome category. Differential errors differ between groups; non-differential errors are not necessarily harmless.

Confounder: Associated with exposure, a **risk factor for outcome**, and **not on the causal pathway**. Smoking can confound an association between coffee and lung cancer.

Mediator / effect modification: An intermediate causal step / a genuinely different effect between subgroups on the chosen measure. Effect modification should be reported, not treated as an error.

Restriction / matching: Limit eligibility to a confounder category / select comparable participants. Matching also needs appropriate analysis.

Stratification / adjustment: Compare within confounder categories / account for measured confounders statistically.

Diagnostic concepts

Screening / reference standard: Testing apparently healthy people for possible disease / best available method for defining true disease status. A positive screen is not a confirmed diagnosis.

Sensitivity / specificity: Positive tests among **diseased** people / negative tests among **non-diseased** people.

PPV / NPV: Disease among **positive tests** / no disease among **negative tests**. Lower prevalence lowers PPV and raises NPV if sensitivity and specificity stay fixed.

Cutoff: If high values indicate disease, lowering the threshold increases sensitivity but decreases specificity; raising it does the reverse.

Randomised trials

RCT: Randomised controlled trial: chance assignment to intervention and comparison groups, followed by outcome comparison.

Randomisation / random sampling: Chance determines treatment group / chance determines who enters the sample. Different purposes.

Random allocation: Prevents systematic allocation differences; balances known and unknown confounders **on average**, not perfectly in every trial.

Allocation concealment: Hide the next assignment from those enrolling people **before allocation**. Prevents manipulation of recruitment or assignment.

Blinding / masking: Keep specified people unaware of treatment assignment **after allocation**. Can reduce differences in behaviour, care and assessment.

Single-blind / double-blind: One party, such as participants or assessors, is unaware / participants and researchers are unaware. State exactly who was blinded.

Control / placebo: Comparator: usual care, another treatment or placebo / a treatment-like intervention without the specific active component.

Protocol / adherence: The planned study procedures / how closely participants follow them.

Intention-to-treat (ITT): Analyse all randomised people in their **original groups**, regardless of adherence, switching or stopping. Preserves the randomised comparison; missing outcomes still require handling.

Per-protocol (PP): Analyse those who followed the protocol. Exclusions can introduce bias because adherence is not random.

As-treated: Group by treatment actually received rather than assigned. Can introduce confounding.

Efficacy / effectiveness: Benefit under ideal or controlled conditions / benefit in practical use.

Example: assigned A but switched to B: ITT keeps the person in A; as-treated puts them in B; PP may exclude them.

Detection and discrimination

Detection changes: More screening or awareness can raise diagnosed incidence without more disease developing. Keep case definitions and detection methods comparable.

Lead-time bias / overdiagnosis: Earlier diagnosis lengthens measured survival without delaying death / finding disease that would never cause symptoms or harm in that person's lifetime.

ROC / AUC: Receiver operating characteristic: sensitivity vertically against **1 − specificity** horizontally. Area under the curve: 0.5 = chance ranking; 1 = perfect discrimination, not accuracy at one cutoff.

Interpreting an association: name the groups, outcome, measure and time period. Say “associated with” unless a causal conclusion is justified. RR/OR alone proves neither causality nor statistical significance.

Prevalence and incidence risk

A fixed population has 200 people, including 20 existing cases. During one year, 9 others develop the condition. Nobody enters or leaves.

Point prevalence = existing cases / population
20/200 = 0.10 = **10%**

Interpret: At baseline, **10 in every 100 people had the condition**. This includes existing cases, not just new ones.

Incidence risk = new cases / initially at-risk people
9/(200 − 20) = 9/180 = 0.05 = **5%**

Interpret: Over one year, **5 in every 100 initially disease-free people developed the condition**. Remove existing cases from this denominator.

Period prevalence = people affected during period / population
(20 + 9)/200 = **14.5%**

Interpret: 14.5% had the condition at some point during the year. This is not necessarily end-of-year prevalence: some may have recovered.

Person-time and incidence rate

Seven people contribute 3, 3, 4, 3, 6, 2 and 6 years at risk; 2 become cases. Count only time before onset, loss, death or study end.

Total = 3+3+4+3+6+2+6 = **27 person-years**
Rate = 2/27 × 100 = **7.41 per 100 person-years**

Interpret: There were **7.41 new cases per 100 person-years of observation at risk**. This is a rate, **not a 7.41% one-year probability**. Person-years combine time across people.

Risk and odds

With 20 cases among 100 initially at-risk people:

Risk = 20/100 = **20%**
Odds = 20/80 = **0.25 = 1:4**
Odds = risk/(1 − risk)
Risk = odds/(1 + odds)

Interpret: 20% became cases; there was **one case for every four non-cases**. Odds of 0.25 do not mean 25% risk.

Birth and mortality rates

In one year, a population of 10,000 has 120 live births and 80 deaths.

Birth rate = 120/10,000 × 1,000 = **12**
Crude death rate = 80/10,000 × 1,000 = **8**

Interpret: There were **12 live births and 8 deaths per 1,000 inhabitants during the year**. For age-specific rates, use only that age group in both numerator and denominator.

Direct age-standardisation

Younger rate = 1 per 1,000; older rate = 20 per 1,000. Standard population: 80% younger and 20% older.

Standardised rate = sum(rate × standard weight)
(1 × 0.80) + (20 × 0.20) = **4.8 per 1,000**

Interpret: With the **standard age distribution**, the combined rate would be 4.8 per 1,000. This enables age-adjusted comparison, not a count of actual observed events. Weights sum to 1.

RR, OR and excess risk

	Disease +	Disease −
Exposed	a = 80	b = 20
Unexposed	c = 40	d = 60

R_E = a/(a+b) = 0.80; R_U = c/(c+d) = 0.40
RR = 0.80/0.40 = **2**
OR = ad/bc = (80 × 60)/(20 × 40) = **6**
Excess risk = 0.80 − 0.40 = **40 percentage points**

Interpret: Exposed people had **twice the risk**, but **six times the odds**, of disease. There were **40 additional cases per 100 exposed people** over the same follow-up. These are different measures.

In a case-control study, OR = (a/c)/(b/d): cases had six times the **odds of exposure** compared with controls. Sampled totals do not give population risks.

Other ratios and attributable fraction

Rate ratio: (20/1,000 person-years)/(10/2,000 person-years) = 4. New cases arose at **four times the rate** in the exposed group.

Prevalence ratio: 12%/6% = 2. Disease was **twice as prevalent** in group A; this does not mean twice the risk of developing it.

Reverse RR: if exposed/unexposed = 4, unexposed/exposed = 1/4 = 0.25: **75% lower risk**, not an absolute risk of 25%.

Attributable fraction: (0.80 − 0.40)/0.80 = 50%. **If causal, half of cases among exposed people are attributable to the exposure.**

Treatment benefit: ARR, RRR and NNT

One-year adverse-outcome risk is **20% with control and 12% with treatment**. CER = control event risk; EER = experimental/treatment event risk.

ARR = CER − EER = 0.20 − 0.12 = **0.08**
RR = EER/CER = 0.12/0.20 = **0.60**
RRR = ARR/CER = 0.08/0.20 = **40%**
NNT = 1/ARR = 1/0.08 = 12.5 → **13**

> **Interpret: ARR:** 8 fewer adverse outcomes per 100 treated rather than control over one year. **RR:** treated risk is 60% of control risk. **RRR:** risk is 40% lower, not 40 percentage points lower.

> **Interpret: NNT:** treating 13 people rather than giving the comparator prevents **one additional adverse outcome on average over one year**. It does not mean that only one person benefits in any way.

Use ARR as a proportion and **round NNT up**. If using percentage points: NNT = 100/8. Never calculate 1/8.

Working backwards from NNT

ARR = 1/NNT as a proportion
ARR = 100/NNT in percentage points
Expected additional benefits = number treated/NNT
Treated adverse risk = control risk − ARR

If the exact NNT is **20** over one year:

Find	Calculate and interpret
ARR	1/20 = 0.05: **5 fewer cases per 100** treated rather than control.
Benefit in 200 treated	200/20 = **10 additional cases prevented on average**, versus control.
Treated risk if control risk = 20%	20% − 5 percentage points = **15%**.
RRR in that example	5/20 = **25% relative reduction**.

NNT alone cannot give treated risk or RRR without control risk. If NNT was rounded, 1/NNT only approximates the original ARR.

Comparing NNT and calculating NNH

Lower NNT = larger absolute benefit, when outcome, population and follow-up are comparable. NNT 20 prevents more events per treated person than NNT 50. If ARR = 0, NNT is infinite: no additional benefit.

NNH = 1/(treated adverse risk − control risk)
Risk 2% → 4%: 1/(0.04 − 0.02) = **50**

> **Interpret:** Treating 50 people causes **one additional adverse event on average** over the stated period. Lower NNH means more frequent additional harm.

Diagnostic calculations and interpretations

	Disease +	Disease −
Test +	TP = 90	FP = 45
Test −	FN = 10	TN = 855

Disease status comes from the reference standard. There are 100 diseased and 900 non-diseased people.

Sensitivity = TP/(TP+FN) = 90/100 = **90%**

> **Interpret: 90% of people with disease test positive.** The remaining 10% are false negatives.

Specificity = TN/(TN+FP) = 855/900 = **95%**

> **Interpret: 95% of people without disease test negative.** This is not the probability of being disease-free after a negative test.

PPV = TP/(TP+FP) = 90/135 = **66.7%**

> **Interpret: 66.7% of people with a positive result have disease.** About one-third of positive results are false positives.

NPV = TN/(TN+FN) = 855/865 = **98.8%**

> **Interpret: 98.8% of people with a negative result do not have disease.** About 1.2% of negative results are false negatives.

False results: do not swap denominators

False-positive rate: FP/(FP+TN) = 45/900 = 5%. Of people **without disease**, 5% test positive. Equals 1 − specificity.

False positives among positive results: FP/(TP+FP) = 45/135 = 33.3%. Of **positive tests**, 33.3% are false. Equals 1 − PPV.

False-negative rate: FN/(FN+TP) = 10/100 = 10%. Of people **with disease**, 10% test negative. Equals 1 − sensitivity.

Build a table from prevalence

With 1,000 people, prevalence 10%, sensitivity 90% and specificity 95%:

Diseased = 1,000 × 0.10 = **100**
Non-diseased = 1,000 − 100 = **900**
TP = 100 × 0.90 = **90**; FN = 100 − 90 = **10**
TN = 900 × 0.95 = **855**; FP = 900 − 855 = **45**

> **Interpret:** Of all 1,000 people, 90 have disease and test positive; 10 have disease but are missed; 855 are disease-free and test negative; 45 are disease-free but test positive.