0 · How To Use This

READ FIRST

★ This subject is a **pipeline**: UNDERSTAND (is there a genetic role? — aggregation, heritability) → DISCOVER (which variants? — LD, GWAS, MR) → CHARACTERISE (how risky? — penetrance, modifiers, G×E) → USE (screening). Side 1 = understand & discover; side 2 = characterise & use.

Assessment shape: online MCQ **10%** (10 Qs, 1-week window) + written A1 **40%** (Modules 1–3) + A2 **50%** (Modules 4–8). All **online / take-home** — no invigilated exam.

Every assignment task is one of three: (a) calculate + interpret, (b) discuss findings, (c) **critically appraise** a design. So the two high-value moves are: plug the right formula, then judge the design's bias. A Stata `.do` file is even handed out for A1 Q1 — expect software-based calculation, then a written interpretation.

SIA → *The mantra that earns marks everywhere: aggregation / high MZ-vs-DZ correlation is "evidence for, but not proof of, an inherited genetic aetiology." Say it whenever you interpret aggregation or heritability.*

1 · Genetics Primer

EXTRA-MODULE 1

Locus = position on a chromosome. **Allele** = the base(s) there; *minor allele* = rarer one. **Genotype** = the pair (e.g. TT, TC, CC); homo- vs hetero-zygous.

- **Polymorphism** — common variant (>1% freq), e.g. a **SNP**; small/no effect
- **Pathogenic mutation** — major deleterious effect → big risk
- **Germline** = inherited, in *every* cell → familial risk (sample blood/buccal)
- **Somatic** = acquired, tumour only → not inherited (sample biopsy)
- **Minor allele** = the less common allele at the locus in the population

Mutation classes: silent (usually benign), **missense** (changes amino acid), **nonsense** (premature stop), **frameshift** indels (corrupt every downstream codon → usually pathogenic). **CNV** = larger gain/loss.

2 · Modes of Inheritance

RISK INEQUALITIES

Defined on *Pr(phenotype | # risk alleles)*, not on "having" the trait:

AUTOSOMAL

```
Dominant: Pr(2)=Pr(1) > Pr(0)
Recessive: Pr(2) > Pr(1)=Pr(0)
Codominant: Pr(2) > Pr(1) > Pr(0)
```

Carrier risk can be **<1** (incomplete penetrance) and non-carrier risk **>0** (phenocopies/sporadic). So a dominant variant can still have penetrance below 100%.

SEGREGATION (PUNNETT)

Each parent passes **one** randomly-chosen allele. **Aa×aa** → ½ Aa, ½ aa (no AA). **Aa×Aa** → ¼ AA, ½ Aa, ¼ aa ⇒ P(child carries ≥1 A)=¾, P(AA)=¼.

Trap: the genotype gives the *expected* probability distribution, not the realised counts in a small sibship.

2b · Germline vs Somatic

SAMPLE CHOICE

- Inherited colorectal-cancer family risk → **germline** → sample blood / buccal swab
- Tumour responds differently to chemo, no family history → **somatic** → sample the tumour biopsy

3 · Allele & Genotype Freq

CALCULATE

From counts n(AA), n(Aa), n(aa) in N people:

ALLELE FREQUENCY (PER CHROMOSOME)

```
p = [2·n(AA) + n(Aa)] / 2N · q = 1 - p
```

Worked: 100 people = 64 CC, 32 CT, 4 TT. T alleles = 2·4+32 = 40; total alleles = 2·100 = 200 ⇒ freq(T)=40/200=0.20, freq(C)=0.80 (20% of all alleles at this locus are T).

CARRIER FREQUENCY (PER PERSON)

```
carrier freq = p² + 2pq = 1 - q²
```

Worked: risk-allele freq 0.1 ⇒ $0.1^2 + 2(0.1)(0.9) = 0.01 + 0.18 = 0.19$ (19% carry ≥1 copy). Equivalently $1 - q^2 = 1 - 0.81 = 0.19$.

Why it matters: the variant is the *exposure*; carrier freq = exposure prevalence → drives sample size/power. Rare variants need huge or enriched samples. **Trap:** allele freq (per-chromosome, ÷2N) ≠ carrier/genotype freq (per-person, ÷N). At T freq 0.01, TT is very rare (q^2=0.0001) yet carriers are ~2% — design power around the *carrier* count.

4 · Hardy–Weinberg

CANON

Holds in a **large, randomly-mating** population with **no selection, migration or mutation** ⇒ genotype freqs are constant across generations & predicted by allele freqs. This is exactly the genotype split used for carrier frequency:

HWE

```
p² + 2pq + q² = 1
AA=p² · Aa=2pq · aa=q²
```

TEST (X^2 GOODNESS-OF-FIT)

```
χ² = Σ (O - E)² / E · df = 1
significant if χ² > 3.84 (α=0.05)
```

df=1: 3 genotype classes − 1 − 1 (estimated allele freq). Deviation in CONTROLS ⇒ genotyping error / population stratification → GWAS QC check.

Worked: p(T)=0.20 in N=100 ⇒ expected 100·0.2²=4 TT, 100·2(0.2)(0.8)=32 TC, 100·0.8²=64 CC. Observed 4/32/64 match exactly ⇒ $\chi^2 \approx 0$ ⇒ in HWE (QC passes). If instead observed = 10 TT, 20 TC, 70 CC (allele freq still ≈0.20), then $\chi^2 = (10-4)^2/4 + (20-32)^2/32 + (70-64)^2/64 \approx 9 + 4.5 + 0.6 =$ **14.1 > 3.84** ⇒ reject HWE ⇒ in controls, suspect a genotyping error or population stratification and exclude/recheck the SNP.

Trap: HWE deviation in *cases* can be a real disease association — so test HWE conventionally in **controls**.

5 · Linkage Disequilibrium

WHY GWAS WORKS

Two loci in **LD** = their genotypes are statistically *correlated* in a random person; nearby loci co-inherited. A marker SNP associated with disease flags a nearby causal variant.

LD MEASURES

```
D = P(AB) - P(A)P(B)
D' = D/D_max ∈ [-1,1] · D'=1 ⇒
complete LD
r² = D² / [P(A)P(a)P(B)P(b)] ∈ [0,1]
```

r^2 is the metric that matters for tagging/power: r^2=1 ⇒ marker perfectly proxies the causal SNP; r^2=0.5 ⇒ need ~2× the cases to detect the same indirect signal. A **haplotype** = the specific alleles inherited together on one chromosome.

Trap: D' and r^2 answer different questions. D'=1 (no recombination) can coexist with *low* r^2 when the two SNPs have different allele frequencies — for tagging/power it is **r^2**, not D', that counts.

6 · Familial Aggregation

MODULE 1

Families share **genes + environment** + can be followed over time. Stronger aggregation in *genetically closer* relatives ⇒ evidence for (not proof of) inherited aetiology — because closer relatives also share more environment.

DEGREE	RELATIVES	GENES SHARED
1st	parents, sibs, children	½
2nd	grandparents, aunts, half-sibs	¼
3rd	first cousins	⅛

DESIGN → MEASURE → BIAS

DESIGN	MEASURE	WATCH
Case-control	**OR**	recall, selection
Retro cohort	RR, SMR	recall, selection
Prospective fam.	RR/HR	slow; **no recall bias**
Twin	heritability	not pop-repr.
Adoption	genes vs env	rare, hard
Migrant	rate compare	healthy-migrant

7 · Aggregation Measures

PLUG NUMBERS

From the 2×2 (proband case/control × relative affected/unaffected), cells a,b,c,d:

EFFECT ESTIMATES

```
OR = (a·d)/(b·c)
RR = [a/(a+b)] / [c/(c+d)]
SMR = Observed / Expected
λ_R = risk in type-R relative /
prevalence K
FRR = RR given affected 1st-degree
relative
```

SMR worked: mothers of cases O=45, E (population rates × person-time) =17.7 ⇒ SMR ≈ 2.5. **λ_R** >1 and *declining* with relatedness ⇒ genetic; the rate of decline hints polygenic vs single-gene. OR ≈ RR only when disease is rare.

OR worked: any affected sister 13/462 in cases vs 1/405 in controls ⇒ OR = (13·404)/(449·1) ≈ 11.7 (95% CI 1.7–98.2). The very wide CI (only one exposed control) ⇒ imprecise — report the CI, not just the point estimate, and beware the small-cell instability.

8 · Migrant & FH Quality

INTERPRET

- Migrant rate stays like **source** ⇒ genetics (or similar env)
- Shifts toward **host** ⇒ environment
- Migrant vs descendants differ ⇒ a **critical age** of exposure

Family-history misclassification: non-differential (random) ⇒ bias toward null; differential (cases recall better) ⇒ bias away from null, inflating OR/RR. Fix with standardised questionnaires, multiple informants, validation against registries/pathology/death records, trained interviewers.

8b · Family Designs

M1 EXTRAS

Case-control-family / case-family: relatives directly interviewed ⇒ OR / RR / **SMR**; relatives of controls are hard to recruit, and the case-family design needs a **population registry**.

Outcome can be analysed as **dichotomous** (affected y/n), **ordinal** (number affected) or **multinomial** — match the analysis to how family history was coded.

9 · Heritability

MODULE 2

= proportion of **phenotypic variance** due to **genetic variance**. A property of a *population in an environment*, not an individual. Variance = SD^2 (e.g. height SD 9.29 ⇒ variance ≈ 86).

VARIANCE PARTITION

```
Vp = Vg + Ve
Vg = Va + Vd (+ Vi)
Broad-sense H² = Vg/Vp
Narrow-sense h² = Va/Vp (h² ≤ H²)
```

Narrow-sense (additive Va) predicts relative resemblance & response to selection; Vd = dominance, Vi = epistatic/interaction variance. Estimate variance **separately by sex & zygosity** (M>F; DZ>MZ spread).

10 · Twin Studies

THE ENGINE

MZ share ~100% genes; **DZ** ~50% (like full sibs). Both share rearing env → comparing them isolates genetics; twins control for age & shared env.

Binary: concordance = proportion of pairs both affected; conc_MZ > conc_DZ ⇒ genetic. **Continuous:** correlate twin-1 vs twin-2.

FALCONER'S HERITABILITY

```
h² = 2 (r_MZ - r_DZ) (continuous)
h² = 2 (conc_MZ - conc_DZ) (binary)
```

Worked: female height r_MZ=0.78, r_DZ=0.46 ⇒ h^2 = 2(0.78−0.46) = 0.64 — 64% of variance in female height is additively genetic. Interpret: "consistent with, but not proof of, an inherited genetic aetiology."

Genetic variance from heritability: Vg = h^2 × Vp. With Vp≈86 and h^2=0.64 ⇒ Vg≈55. Opposite-sex DZ pairs & the twin–co-twin (TRA) design extend the model to probe shared-environment and sex effects.

11 · ACE Model

VARIANCE COMPONENTS

Split Vp into **A** additive genetic, **C** common/shared env, **E** unique env + error. From twin correlations:

ACE FROM R_MZ, R_DZ

```
r_MZ = A + C · r_DZ = ½A + C
A = 2(r_MZ - r_DZ) (= Falconer)
C = 2·r_DZ - r_MZ · E = 1 - r_MZ
```

Worked: r_MZ=0.78, r_DZ=0.46 ⇒ A=2(0.78−0.46)= 0.64; C=2(0.46)−0.78=0.14; E=1−0.78=0.22. Check: A+C+E = 0.64+0.14+0.22 = 1.00 ✓.

So C is the part of resemblance shared equally by both twin types; E (incl. measurement error) is the only thing that makes MZ co-twins differ. **Trap — equal-environments assumption:** if MZ pairs are treated more alike than DZ, shared env masquerades as genes ⇒ h^2 overestimated.

11b · Classic Twin Model

4 ASSUMPTIONS

- MZ share A=1.0, DZ share A=0.5 (like full sibs)
- MZ & DZ share C **equally** (equal-environments)
- Random mating (no assortative mating inflating r_DZ)
- No gene–environment interaction/correlation
- Trait measured the same way in both twin types

Break any assumption ⇒ biased h^2.

Concordance/correlation are estimated **separately by sex & zygosity** because variance differs.

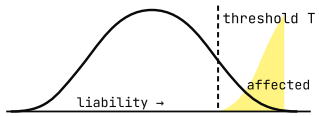
Binary worked: conc_MZ=0.40, conc_DZ=0.15 ⇒ h^2(liability) = 2(0.40−0.15) = 0.50. MZ>DZ concordance is the signal; near-equality (conc_MZ≈conc_DZ) ⇒ shared environment, not genes, drives the resemblance.

12 · Liability-Threshold

BINARY TRAITS

Assume an unobserved continuous **liability** (genes+env), ~Normal; disease occurs above a **threshold** set by prevalence. Puts yes/no disease onto a continuous scale so variance/heritability methods apply.

LIABILITY ~NORMAL · DISEASE = TAIL BEYOND THRESHOLD



Tail area = prevalence. Relatives of cases sit at a *right-shifted* liability distribution ⇒ larger tail ⇒ higher risk, the model's link from heritability to a yes/no trait.

Trap: heritability of *liability* ≠ heritability "of the disease," and is very sensitive to the assumed prevalence (which sets where T sits).

13 · Heritability Cautions

ASSIGNMENT GOLD

High h^2 does **NOT** mean: (a) the trait is unmodifiable; (b) genes cause *between-group*/between-population differences; or (c) anything about an *individual*. It is a population- & environment-specific quantity.

Missing heritability: GWAS-discovered SNPs explain far less variance than the twin-study h^2. Candidate causes: private (family-specific) mutations, rare moderate-risk variants, additional undiscovered common SNPs, gene–gene interactions, and non-genetic factors correlated within relatives.

So twin-estimated h^2 and GWAS-explained variance are different quantities — don't expect the discovered SNPs to "add up" to the twin h^2. High h^2 ≠ "untreatable": environment can still shift the whole distribution (height is highly heritable yet population mean rose with nutrition).

14 · Genetic Association

MODULE 3

= a **case-control study where the exposure is a genetic marker** (a SNP). Association arises if the SNP *causes* disease, is in *LD* with a causal variant, or is **confounded by ancestry** (stratification).

Candidate-gene = a few pre-specified, biologically-motivated SNPs; **GWAS** = hundreds of thousands–millions of SNPs, scanned agnostically across the whole genome. The marker is the *exposure*; cases vs controls are compared on marker frequency, reported as an OR + 95% CI per SNP.

An association is **useful for prediction even if non-causal**. Three reasons a SNP associates with disease:

- the SNP **causes** disease (directly functional)
- it is in **LD** with a nearby causal variant (still useful for prediction)
- **artefact** of confounding by ancestry (stratification)

Only the first two replicate in an independent sample — the third is what replication + PC-adjustment are designed to kill.

A genetic/polygenic **risk score** sums many such SNPs and is ~Normal in the population, sliding people along a continuous risk axis rather than a single yes/no genotype — the basis for risk stratification in M8.

15 · Association Tests

X^2 / LOGISTIC

TEST	TABLE	DF
Allelic	2×2 allele×status	1
Genotypic	2×3 genotype×status	2
Dominant	AA+Aa vs aa	1
Recessive	AA vs Aa+aa	1
Additive	per-allele 0/1/2	1

CHI-SQUARE & LOGISTIC OR

```
χ² = Σ(O-E)²/E → large χ² → small p
logit P(D) = β₀ + β₁·genotype +
covariates
OR = e^β₁ · OR = (a·d)/(b·c)
```

Per-allele coding (0,1,2) ⇒ OR per extra risk allele; adjust for ancestry **principal components**, age, sex. State the mode of inheritance up front; testing several models multiplies the tests and so needs a stricter threshold.

16 · Multiple Testing

THE GWAS PROBLEM

Testing millions of SNPs hugely inflates the **type-1 error / false-positive** rate; at α=0.05, 1 in 20 truly-null SNPs looks "significant" by chance alone.

THRESHOLDS

```
Bonferroni: α = 0.05 / (# tests)
genome-wide significance = 5×10⁻⁸
```

$5\times10^{-8} \approx 0.05 / 10^6$ independent common-variant tests; hits must replicate independently. **Worked:** a candidate study of 50 SNPs ⇒ Bonferroni α = 0.05/50 = 0.001 — a SNP at p=0.01 is *not* significant after correction. **Trap:** Bonferroni is conservative (LD makes tests correlated) but 5×10^{-8} is the field standard — use it for GWAS.

17 · Manhattan & QQ

READ THE PLOT

Manhattan: x = genomic position, y = $-\log_{10}$(p). Peaks crossing $-\log_{10}(5\times10^{-8}) \approx 7.3$ = associated loci.

QQ plot: observed vs expected $-\log_{10}$(p) under the null. On the diagonal = no inflation; an *early, whole-line* upward lift = stratification / cryptic relatedness / artefact (genomic inflation λ_GC; λ≈1 is good); a departure only in the *extreme tail* = genuine signal.

Trap: don't read a single Manhattan peak as "the causal gene" — the top SNP is usually the best *tag* in LD with the true causal variant, so fine-mapping is needed to localise the cause.

18 · Pop. Stratification

KEY CONFOUNDER

Cases & controls differ in **ancestry**; both allele freqs & disease rates vary by ancestry ⇒ spurious association (confounding). **Fixes:** match on ancestry, adjust for **principal components**, genomic control (λ_GC), or family-based designs; HWE deviation in controls helps flag it.

This is why a hit must **replicate in an independent sample** and why GWAS report λ_GC — a clean QQ plot (λ≈1) is the reassurance that genuine signal, not stratification, is driving the Manhattan peaks. λ > 1 ⇒ inflate-corrected before trusting any hit.

Formula Belt

SIDE 1

```
p=[2n(AA)+n(Aa)]/2N · carrier=p²+2pq
HWE p²+2pq+q²=1 · χ²=Σ(O-E)²/E df1
r²=D²/[P(A)P(a)P(B)P(b)] · OR=ad/bc
h²=2(r_MZ-r_DZ) · A=2(r_MZ-r_DZ)
SMR=O/E · λ_R=relative risk/K · GWS
5×10⁻⁸
```

19 · Mendelian Randomisation

MODULE 4

Use a **genetic variant as an instrumental variable (IV/proxy)** for a modifiable exposure to test *causation*. Genotype is randomly allocated at conception ("nature's RCT") ⇒ not subject to reverse causation or conventional confounding.

It mimics an RCT's randomisation: alleles are dealt independently of the lifestyle/environmental confounders that wreck observational X–Y comparisons, and a fixed germline genotype *can't* be changed by the disease (no reverse causation). The question it answers is *"does X cause Y,"* using a variant that proxies lifelong X.

20 · The 3 IV Assumptions

STATE VERBATIM

1. **Relevance** — the proxy is *robustly* associated with the exposure (must be strong for adequate power)
2. **Independence** (exchangeability) — proxy independent of confounders of the X–Y relationship
3. **Exclusion restriction** — proxy affects the outcome *only* through the exposure (no direct or alternative path)

DAG & WALD ESTIMATE

```
G → X → Y (G ⊥ U; no direct G → Y)
β(X→Y) = β(G→Y) / β(G→X)
```

If G associates with Y and all 3 hold, X *likely* causes Y — MR sits on a **continuum convincing → not**; assumptions are argued *likely*, never proven. Assumption 1 is testable (the G–X association); 2 and 3 are largely untestable and argued from biology, so MR conclusions are framed as supporting (not proving) a causal role.

21 · MR Threats

APPRAISAL TARGETS

- **Horizontal pleiotropy** — variant affects Y via another pathway ⇒ breaks exclusion (the #1 threat); probe with MR-Egger, weighted median
- **Weak instrument** — breaks relevance ⇒ low power, bias toward the confounded observational estimate
- **Confounding via LD / stratification** — instrument correlated with another causal variant
- **Canalisation** — developmental compensation; lifelong genetic exposure ≠ a short intervention

Trap: MR estimates a *lifelong average* effect — answers "does X cause Y," not "what if I change X for 6 months." Course examples: insulin-resistance gene scores → renal/pancreatic cancer; vitamin-B12 genes → lung cancer (supports a causal role).

21b · Wald Ratio Worked

SHOW THE NUMBER

The ratio (Wald) estimate divides the variant–outcome effect by the variant–exposure effect. Say G raises the exposure by β(G→X)=0.5 units per allele, and G is associated with the outcome at β(G→Y)=0.1 (log-odds per allele):

```
β(X→Y) = 0.1 / 0.5 = 0.2 per unit of X
```

Interpretation: each one-unit higher (genetically-predicted) exposure ⇒ 0.2 higher log-odds of disease — a *causal* estimate **if** the 3 assumptions hold. A weak instrument (small β(G→X)) blows up the ratio's variance ⇒ check the F-statistic. Combine many SNPs by inverse-variance weighting; MR-Egger & weighted-median are the robustness checks for pleiotropy.

22 · Penetrance

MODULE 5

= **probability of disease by a specific age (or over a period)** for a person with a given genotype, possibly conditional on covariates. E.g. *MSH6* variant → colorectal-cancer penetrance ≈ 50% by age 70 (males).

Complete = all carriers eventually affected; **incomplete** = penetrance <1 (most disease genes).
Age-specific / cumulative = a curve of cumulative risk vs age, typically by **survival analysis / Kaplan-Meier** birth→diagnosis.
Expressivity (contrast): penetrance = *whether* disease occurs; variable expressivity = *how severe / which features*. Penetrance may also be reported **by sex** and conditional on covariates, and is the input to risk-based counselling.

23 · Estimating Penetrance

DESIGN-SPECIFIC

DESIGN	HOW	NEEDS
Case-control	OR → absolute risk	population incidence
Prosp. cohort	follow carriers → survival	large N (rare)
Family / weighted	clinic carriers + weights	registry rates

Case-control gives OR; convert to absolute (age-specific) risk using **non-carrier / population incidence** — penetrance needs external incidence data.
Prospective carrier cohorts need **large N** because high-risk variants are rare ⇒ low power. **Trap — ascertainment bias:** clinic carriers are tested *because* of strong FH / young onset ⇒ not random ⇒ naïve estimates overestimate penetrance.

24 · Weighted Cohort

MODULE 6 · SIGNATURE

Fix for non-random ascertainment — build a **"synthetic cohort"** mimicking carriers drawn randomly from the population by probability weighting:

1. Age/sex carrier incidence = **population incidence × RR for carriers**
2. Derive **weights** so affected:unaffected per age-stratum matches population proportions
3. Analyse weighted data ⇒ unbiased, generalisable penetrance

Also called **modified segregation analysis** when carrier status is inferred across the family rather than directly genotyped.

25 · Modifiers of Penetrance

MODULE 6

Genetic/environmental factors that **alter risk among carriers of the same variant** — explaining why same-gene carriers span "modest" to "extreme" risk, not clustered at the average. Use for pathogenesis, risk reduction, **individualised counselling + risk-based screening**.
Trap: a modifier acts *within carriers* — distinct from a general-population main effect and from whole-population G×E (M7).
Modifiers explain the **wide spread** of carrier risk; the same weighted-cohort machinery (M6) estimates a modifier's effect by re-weighting clinic-ascertained carriers to a synthetic random cohort, then comparing risk across modifier strata. Output → risk-stratified screening & counselling.

26 · Gene–Environment Interaction

MODULE 7

G×E exists when the **exposure–disease association differs across genotypes** (equivalently, the genotype effect differs across exposure levels). Statistical interaction = a departure from a *specified* no-interaction model ⇒ it is scale-dependent.

NO INTERACTION MEANS…

```
Multiplicative: RR_joint = RR_G × RR_E
Additive: RD_joint = RD_G + RD_E
(RERI=0)
```

Multiplicative is the **default output** of logistic/Cox models (they multiply ORs/HRs); additive needs the absolute risk differences. **Synergistic** = joint effect bigger than expected; **antagonistic** = smaller — interpreted against the underlying biological pathways (shared vs independent mechanisms).

27 · The Classic Trap

STATE THE SCALE

Worked. Disease risk by genotype × exposure:

GENOTYPE	E−	E+	RR	RD
Gene −	0.02	0.04	2.0	0.02
Gene +	0.03	0.06	2.0	0.03

RRs equal (2.0=2.0) ⇒ **NO multiplicative interaction**; RDs differ (0.03≠0.02) ⇒ additive interaction present. Same data, two answers — always state the scale.
Additive (RD) is the public-health-relevant one (who gains most from removing the exposure); multiplicative is the default logistic/Cox output. Synergistic = bigger than expected; antagonistic = smaller. Check the joint cell (gene+/E+ = 0.06) against both the product and the sum.

28 · G×E Designs

DETECT IT

- **Case-control** (standard) — include **G, E and the G×E product** term in logistic regression
- **Case-only** — cases only; test whether G and E are associated *among cases*. Under *G ⊥ E in the source population*, a G–E association estimates the **multiplicative** interaction efficiently (no controls)
- **Cohort / family** designs also detect G×E and G×G, with more power for rare exposures but higher cost

Trap: case-only is biased if G and E are correlated in the population and estimates *only* multiplicative interaction.
Implication: precision prevention — target modifiable environmental exposures in the genetically susceptible.

28b · The Other Case

MULTIPLICATIVE PRESENT

Contrast: if gene+ gave (E−=0.03, E+=0.12) ⇒ RR=4.0 while gene− RR=2.0 ⇒ RRs differ (4≠2) ⇒ multiplicative interaction present. Read RR-*ratio* across strata for multiplicative, RD-*difference* for additive — same data, two verdicts.
RERI (relative excess risk due to interaction) = $RR_{11} - RR_{10} - RR_{01} + 1$; RERI=0 ⇒ no additive interaction, >0 ⇒ synergy, <0 ⇒ antagonism.
G×G (epistasis) is tested the same way — a gene–gene product term — and is one explanation for missing heritability. **Report rule:** always state the scale, give the RR-ratio (multiplicative) *and* the RD-difference (additive), then say which is relevant to the question (public-health ⇒ additive).

29 · Screening

MODULE 8

Disease screening = a systematic test to find asymptomatic disease/precursors in people not seeking care. **Genetic screening** = find risk-raising variants in asymptomatic people so risk can be reduced/prevented; can be population-wide but is usually targeted to high-prior-risk groups (e.g. strong family history).
Course twist: a genetic test can be **"once-and-for-all"** (your germline doesn't change), unlike repeated disease screening. Two uses: screen *for* genetic risk, or use a genetic factor *to screen* for disease.

30 · Wilson–Jungner

WHO 1968

The screening-evaluation checklist (the course adapts all 10 to genetics):

- **Condition** — important problem, recognisable latent/early stage, understood natural history
- **Test** — suitable, acceptable, accurate
- **Treatment** — accepted risk-reduction, agreed policy on whom to treat
- **Facilities** for diagnosis & treatment exist
- **Cost** economically balanced; case-finding a **continuing process**
- Agreed **natural history** & an agreed definition of who counts as a "case"

Trap: "we *can* test" ≠ "we *should* screen." A test only helps if knowing the result reduces disease/disability/death and benefits beat harms (psychological, social, insurance, variants of unknown significance, false positives). Example genes the course uses: *BRCA1/2*, the mismatch-repair (MMR) genes, *HTT* — note *HTT* (Huntington) has no risk-reduction, which weakens the case for screening.

31 · NNT & NNS

QUANTIFY BENEFIT

TWO-STEP SCREENING CALC

```
ARR = carrier risk × proportion risk reduced
NNT = 1 / ARR
NNS = NNT / carrier frequency
```

Worked (BRCA1/2): carrier breast-cancer risk to 70 = 0.4; tamoxifen cuts risk 50% ⇒ ARR = 0.4×0.5 = 0.2 ⇒ NNT = 1/0.2 = 5 carriers treated to prevent one cancer.
Carrier freq 0.0067 (1 in 150) ⇒ NNS = 5/0.0067 ≈ **746** screened per cancer prevented. High-FH group (carrier freq 0.25) ⇒ NNS = 5/0.25 = 20 — far more efficient ⇒ justifies **targeted** screening.

CARRIER FREQ	NNT	NNS
0.0067 (general)	5	≈746
0.05 (moderate FH)	5	100
0.25 (strong FH)	5	**20**

Trap: NNS collapses for rare variants in the general population — raising the prior probability of carriage (targeting high-FH groups) is what makes genetic screening worthwhile.

31b · Harms Ledger

BENEFITS VS COSTS

Screening is only justified when benefit beats harm. Weigh against the NNT/NNS benefit:

- **False positives** → anxiety, over-treatment
- **Variants of unknown significance** → uninterpretable results
- **Psychosocial** → family, identity, fatalism
- **Insurance / legal / discrimination** risk
- **Opportunity cost** of the screening budget

32 · Test Performance

2×2 METRICS

From a 2×2 of test (+/−) × true status (D+/D−): TP, FP, FN, TN.

ACCURACY METRICS

```
Sensitivity = TP/(TP+FN) P(+|disease)
Specificity = TN/(TN+FP) P(-|healthy)
PPV = TP/(TP+FP) · NPV = TN/(TN+FN)
```

Sens & spec are intrinsic to the test; PPV rises & NPV falls as prevalence rises. In low-prevalence screening even a very specific test gives many false positives → low PPV.

BAYES FORM

```
PPV = (Sens·Prev) /
  [Sens·Prev + (1-Spec)(1-Prev)]
```

33 · ROC & AUC

DISCRIMINATION

ROC: plot sensitivity (y) vs 1−specificity (x) as the cut-off moves. **AUC** = P(a random case scores higher than a random control): 0.5 = chance (the diagonal), 1.0 = perfect (top-left corner). In this course AUC appears for **polygenic risk scores** (e.g. coronary-artery-disease AUC ≈ 0.81).
Moving the threshold **trades sensitivity against specificity**. **Trap:** excellent sens/spec is useless for screening if prevalence is tiny (PPV near zero) — always tie performance back to prevalence / carrier frequency.

33b · PPV Worked

PREVALENCE BITES

Sens=0.90, Spec=0.99. At **prevalence 1%**:

```
PPV = (0.9·0.01) /
  [0.9·0.01 + 0.01·0.99] =
0.009/0.0189 ≈ 48%
```

Half of positives are false — despite 99% specificity. At prevalence 10% the same test gives PPV ≈ **91%**.
Lesson: raise the prior (target high-risk) before screening, or most positives are false alarms.
Sens/spec are fixed properties of the test; only PPV/NPV move with prevalence — that single fact answers most "evaluate this screening test" questions. NPV is near-perfect when disease is rare (almost all test-negatives really are well), which is little comfort if the few positives are mostly false.

34 · Risk Reclassification

PRECISION PREVENTION

Adding a genetic factor (e.g. a polygenic score) **re-classifies** individuals across an actionable risk threshold — some move up (newly flagged high-risk), some down (reassured). The value of genetic screening = how many it correctly reclassifies + NNS/NNT, *not* "we can test, so we should." Ties back to the Wilson–Jungner conditions (penetrance understood, accurate test, early actionable stage).
A reclassification is only worthwhile if a person moving above the threshold gains an **effective action** (screening, prophylaxis, risk-reducing surgery). Reclassifying with no actionable consequence adds anxiety without benefit.

34b · Disease Screening

USING GENETICS

Two distinct goals: (1) **screen for genetic risk** in the well → reduce future risk; (2) **use a genetic factor to triage disease screening** — e.g. start colonoscopy earlier / more often in MMR carriers. Both still demand an accurate test + an effective downstream action + favourable NNS in the targeted group.

35 · Appraisal Checklist

LO5 · EVERY MODULE

LO5 (appraisal) threads through *every* module. For any study, answer **design → measure → strength → limitation → bias**:

- **Design?** case-control / cohort / twin / GWAS / MR / family / weighted / case-only / screening
- **Measure?** OR vs RR/SMR/HR; h^2; per-allele OR; Wald β; penetrance; NNT/NNS
- **Confounding?** shared environment (aggregation), ancestry/stratification (GWAS), pleiotropy (MR)
- **Selection?** ascertainment (penetrance), healthy-migrant, control recruitment
- **Information bias?** family-history recall (differential vs non-differential), misclassification direction
- **Power?** rare variant / weak instrument / low r^2
- **Generalisability?** twins, clinic families, ancestry of the GWAS sample
- **Causation?** aggregation/association ≠ cause; MR / replication / dose-response strengthen it
- **Precision/CI?** a wide 95% CI (few exposed) = imprecise — don't over-read a point estimate

BIAS	DIRECTION
Non-differential misclass.	toward null
Differential recall (cases)	away from null
Clinic ascertainment	overestimates penetrance
Weak instrument (MR)	toward observational

SIA → *Marks come from naming the direction of each bias (toward vs away from the null), not just listing it. State the rival explanation, then how the design does (or fails to) rule it out.*

36 · Interpretation Hooks

USE THESE PHRASINGS

- Aggregation / MZ>DZ = "evidence for, *not proof of*, inherited aetiology"
- Heritability is a **population-in-an-environment** property; no individual/between-group claim
- An associated SNP is usually a **tag in LD**; r^2 governs power
- GWS = 5×10^{-8} + independent replication
- MR: relevance, independence, exclusion; chief threat = pleiotropy; conclusions *likely*
- Clinic cohorts **overestimate penetrance** without probability weighting (synthetic cohort)
- Interaction is **scale-dependent** — state additive vs multiplicative
- **PPV depends on prevalence** ⇒ screen the targeted high-risk
- OR ≈ RR **only when disease is rare**; report the 95% CI, not just the point estimate
- "We *can* test" ≠ "we *should* screen" — needs an effective downstream action

Calculation Belt

SIDE 2

```
β(X→Y) = β(G→Y)/β(G→X) (Wald)
ARR = carrier risk × % reduced · NNT = 1/ARR
NNS = NNT / carrier freq
Sens = TP/(TP+FN) · Spec = TN/(TN+FP)
PPV = TP/(TP+FP) · NPV = TN/(TN+FN)
multiplic RR_J=RR_G·RR_E · additive RD_J=RD_G+RD_E
```

SIA → *Show the working: in this subject the marks live in the setup and the interpretation, not the final digit. Always write the formula, the substitution, then one sentence of meaning.*