Introduksjon til analyse av gen-gen og gen-miljø interaksjoner i case-kontroll studier

Lars C. Stene[1,2]

[1] *Divisjon for epidemiologi, Nasjonalt folkehelseinstitutt, Postboks 4404 Nydalen, 0403 Oslo*
[2] *Diabetesforskningssenteret, Aker og Ullevål sykehus, Ullevål sykehus, 0407 Oslo*

Korrespondanse til: Lars C. Stene, Divisjon for epidemiologi, Nasjonalt folkehelseinstitutt, Postboks 4404 Nydalen, 0403 Oslo.
Telefon: 22 04 23 99 Telefaks: 22 04 23 51 E-post: lars.christian.stene@folkehelsa.no

SAMMENDRAG

Hensikten med denne artikkelen er å gi en enkel introduksjon til analyse av gen-gen og gen-miljø interaksjoner med case-kontrolldesign. Hovedmålgruppen er epidemiologer med liten eller ingen erfaring med studier av molekylærgenetiske markører. Grunnleggende begreper som gen, allel, eksponering og interaksjon forklares i lys av dose-responssammenhenger. Prinsipper for såkalt 'case-only' analyse gjennomgås. Det er lagt vekt på bruk av logistisk regresjonsanalyser og sammenhengen med tradisjonell tabellanalyse.

Stene LC. **An introduction to analysis of gene-gene and gene-environment interactions in case-control studies.** *Nor J Epidemiol* 2002; **12** (2): 109-117.

ENGLISH SUMMARY

The objective of this paper is to present a simple introduction to the analysis of gene-gene and gene-environment interactions with the case-control design. The primary target group is epidemiologists with little or no experience in studies of molecular genetic markers. Fundamental concepts such as gene, allele, exposure and interaction are explained in the context of dose-response relationships. Principles for the so-called 'case-only' analysis are presented. Emphasis is put on the use of logistic regression analyses and how these are related to traditional analysis of contingency tables.

INTRODUKSJON

Studier av interaksjoner mellom gener og miljø eller mellom ulike gener er en viktig målsetning i genetisk epidemiologi (1). Hensikten med denne artikkelen er å presentere en enkel introduksjon til analyse av gen-gen og gen-miljø interaksjoner i case-kontroll studier samt såkalte case-only studier, med spesiell vekt på bruk av logistisk regresjon. Jeg vil kort forklare hva som menes med interaksjon i denne sammenhengen. Jeg diskuterer også noen prinsipper for modellering av 'gen-dose respons'. Mange begreper som brukes i genetisk epidemiologi er ikke allment brukt blant epidemiologer, og begrepsforvirringen kan være frustrerende for mange som forsøker å tilnærme seg feltet. For en generell introduksjon til prinsipper og termer i genetisk epidemiologi, se f.eks. Elston (2) og artiklene av Ellsworth & Manolio (3-5). For generelle oversikter over metoder for analyse av gen-miljøinteraksjoner med ulike design, se artiklene av Yang & Khoury og Andrieu & Goldstein (6,7). Denne artikkelen kan også være en nyttig introduksjon til logistisk regresjon for genetikere som har erfaring med tradisjonell tabellanalyse. Grunnleggende introduksjon til logistisk regresjon er beskrevet mange steder, for eksempel i Kleinbaum et al. (8).

HVA ER ET GEN OG HVA ER EKSPONERING?

For folk flest er den nøyaktige betydningen av *gen* eller *genotype* uklar, og begrepene brukes ofte om ulike ting. En tradisjonell, men noe problematisk definisjon av et gen er "en sekvens DNA som koder for et protein". Med denne definisjonen korresponderer gen til locus (flertall: loci), som refererer til posisjon på kromosomet der genet finnes. Store deler av DNA koder ikke for proteiner, men kan allikevel inneholde viktig informasjon. Et locus kan derfor referere til en posisjon på kromosomet som ikke koder for proteiner og derfor strengt tatt ikke er 'gener'. Når det humane genom nå sies å være kartlagt betyr det at sekvensene er kjent i rimelig detalj, men det betyr ikke at man kjenner alle genenes posisjon eller funksjon.

Mange gener finnes i to eller flere varianter som kalles alleler. Et locus hvor det finnes to eller flere ulike alleler i befolkningen kalles et polymorft sete. Enkelte foretrekker å bruke begrepet gen om et allel (1). Når man snakker om 'risikogener' menes som regel risikoalleler. Mennesket har to sett med såkalte homologe kromosomer, og dermed to utgaver av hvert 'gen'. Hvis det finnes flere varianter (alleler) av et gen i befolkningen, kan et individ ha null, ett eller to kopier av en gitt variant. Genotypen (med hensyn på

ett locus) beskrives av de to allelene et individ har i et gitt locus (man kan også beskrive genotyper med hensyn på flere loci). Variasjoner i DNA-sekvenser mellom individer forekommer ofte for deler av DNA som ikke har noen kjent funksjon eller som ikke koder for proteiner. Hvis man med laboratoriemetoder kan bestemme hvem som har de ulike variantene av en gitt DNA-sekvens kan dette benyttes som genetiske markører (se senere). Individer med to kopier av et gitt allel sies å være homozygote med hensyn på dette allelet og individer som bærer én kopi sies å være heterozygote.

I epidemiologi snakker man om eksponering for mulige risikofaktorer. Man kan i en viss forstand også snakke om 'eksponering' for gener, fordi ulike varianter kan føre til 'eksponering' for proteiner med ulik funksjon eller grad av funksjon. Det er derfor ofte naturlig å modellere effekten av gener med tre nivåer: 0, 1 eller 2 kopier av allelet man er interessert i. Ofte består funksjonelle molekyler, slik som HLA (human leukocyte antigen) molekyler, av flere proteinsubenheter som kodes for av ulike gener (loci). I slike tilfeller kan det være naturlig å definere eksponering (genotypen) med hensyn på alle de loci som koder for et funksjonelt molekyl. I de fleste tilfeller må man imidlertid forholde seg til genetiske markører. Genetiske markører er loci med minst to varianter som man i de fleste tilfeller ikke kjenner funksjonen til. Det er med andre ord slik at ikke alle genetiske variasjoner medfører (biologisk) funksjonell forandring. Varianter (alleler) av en genetisk markør kan være assosiert med spesifikke alleler i et annet locus på samme kromosom. Slik assosiasjon mellom alleler i ulike loci kalles koblingsulikevekt (linkage disequilibrium). Flere faktorer påvirker graden av koblingsulikevekt, men koblingsulikevekt forekommer ofte mellom alleler i loci som ligger nær hverandre på kromosomet (1). Det er alltid en mulighet for at en assosiasjon mellom en genetisk markør og en sykdom skyldes koblingsulikevekt mellom varianter av den genetiske markøren og et annet risikoallel på samme kromosom som er kausalt forbundet med sykdom. Dette er et eksempel på confounding.

HVA ER INTERAKSJON?

De fleste vanlige sykdommer er såkalte komplekse sykdommer hvor vi i liten grad kjenner årsakene. Komplekse sykdommer er definert ved at de forårsakes av både flere ulike gener og flere ulike miljøfaktorer. Man sier ofte, noe upresist, at de er sykdommer som oppstår i genetisk sårbare individer, eller at de skyldes en interaksjon mellom gener og miljø (9-12). I en viss forstand er det rimelig å si at de fleste sykdommer er et resultat av både gener og miljø, men i denne artikkelen skal vi diskutere hvordan man kvantitativt kan estimere om og hvordan 'effekten' av et gen er avhengig av tilstedeværelse (eller 'dose') av et annet gen eller en miljøfaktor.

I epidemiologien ble man tidlig klar over at 'kvantitativ interaksjon' er avhengig av hvilken skala man måler effektene på, og at man i visse situasjoner kan komme til ulike konklusjoner om interaksjon avhengig om man estimerer effektene med relativ risiko (multiplikativ risikomodell) eller risikodifferens (additiv risikomodell) (13-15). Selv om opprinnelsen er noe uklar, har enkelte, basert på tradisjonell kvantitativ genetikk antatt at en multiplikativ sammenheng mellom effekten av to gener betyr at de to genene virker sammen i samme biokjemiske (kausale) virkningsmekanisme, mens additive sammenhenger indikerer 'uavhengige' mekanismer (16). En slik tolkning følger også fra probabilistisk uavhengighet mellom to sykdomsårsaker for sjeldne sykdommer (17) og er konsistent med Rothmans kausalmodell (18), mens andre har kritisert dette og hevder at det er vanskelig eller umulig å si noe generelt om biologiske mekanismer basert på observert 'kvantitativ interaksjon' (14,19,20). Det er viktig å være klar over at denne skala-avhengigheten gjelder like fullt i tradisjonell stratifisert tabellanalyse ($2 \times 2 \times 2$ tabeller) og for gen-gen interaksjoner, noe som fortsatt ikke er særlig godt kjent blant genetikere (21). Det at interaksjon er et skala- eller modellavhengig begrep har ført til at enkelte foretrekker å snakke om samlet effekt av to faktorer ('joint effect') sammenlignet med effekten av hver enkelt faktor alene. På den måten kan man betrakte interaksjon i et 'dose-respons' perspektiv (se senere).

Med case-kontroll design kan man estimere oddsratio (relativ risiko), men ikke risikodifferens. I praksis er man derfor som regel bundet til en multiplikativ risiko (odds) modell for fravær av interaksjon, som i praksis betyr fravær av interaksjon hvis den relative risikoen (oddsratioen) forbundet med en faktor er lik ved ulike nivåer av den andre faktoren. Dette er den vanligst brukte modellen i epidemiologi (22). Tabell 1 og tabell 2 viser to ekvivalente, men konseptuelt forskjellige måter å se på interaksjon i case-kontroll design (13,23,24).

Et mål på interaksjonseffekten kan defineres ved $OR_{ED|G=1}/OR_{ED|G=0}$, som er ratioen mellom de to oddsratioene i tabell 1. Den vertikale streken leses som 'gitt', og betyr at man betinger på det som står etter den. Notasjonen her følger prinsippene til Kleinbaum et al. (23), der $OR_{ED|G=1}$ betyr oddsratioen for assosiasjonen mellom eksponering (E) og sykdom (D), gitt at man er 'eksponert for' det aktuelle genet (G=1). Med oppsettet i tabell 2 er det lettere å se at interaksjon kan betraktes i forhold til samlet effekt, eller dose-respons sammenheng. Hvis man foretrekker oppsettet i tabell 2 kan man ekvivalent definere størrelsen på interaksjonseffekten som $OR_{11}/(OR_{10} \times OR_{01})$. Størrelsen på interaksjonseffekten, definert på denne måten kalles av og til theta (θ): $\theta = OR_{11}/(OR_{10} \times OR_{01}) = OR_{ED|G=1}/OR_{ED|G=0}$. Det følger at $\theta = 1$ betyr ingen interaksjon i den multiplikative risikomodellen. $\theta > 1$ indikerer positiv interaksjon (synergisme) og $\theta < 1$ indikerer negativ interaksjon (antagonisme). Vi skal se

Tabell 1. 2 × 2 tabell stratifisert i henhold til risikogenotype (2 × 2 × 2 tabell)*.

	Risikogenotype (G=1)	
	Eksponert	Ikke eksponert
Kasus	*a*	*b*
Kontroller	*c*	*d*

$OR_{ED|G=1} = ad/bc$

	Normal genotype (G=0)	
	Eksponert	Ikke eksponert
Kasus	*e*	*f*
Kontroller	*g*	*h*

$OR_{ED|G=0} = eh/fg$

Fravær av interaksjon i en multiplikativ risiko (odds) modell hvis:

$$OR_{ED|G=1} = OR_{ED|G=0}$$

Definerer θ som mål på interaksjon (i multiplikativ modell):

$$\theta = OR_{ED|G=1}/OR_{ED|G=0} = (ad/bc)/(eh/fg) = adfg/bceh$$

* Risikogenotype (G) er definert som en dikotom variabel (G kodet 1 hvis ja og 0 hvis nei). Eksponering (E) er definert som en dikotom variabel (E kodet 1 hvis ja og 0 hvis nei).

ED indikerer at oddsratioen refererer til assosiasjonen mellom eksponering og sykdom (disease), OR_{ED}. Den vertikale streken leses som 'gitt at', slik at $OR_{ED|G=1}$ betyr oddsratioen for assosiasjonen mellom eksponering og sykdom gitt at man har risikoallelet (G=1).

Tabell 2. Tabelloppsett 1 for analyse av interaksjon mellom gener og miljøeksponering i case-kontroll studie (2 × 4 tabell)*.

Eksponering (E)	Risiko genotype (G)	Antall cases	Antall kontroller	Oddsratio
1	1	*a*	*c*	$OR_{11} = ah/cf$
1	0	*b*	*g*	$OR_{10} = bh/fg$
0	1	*e*	*d*	$OR_{01} = eh/df$
0	0	*f*	*h*	$OR_{00} = 1.0$ (ref.)

Fravær av interaksjon i multiplikativ risiko (odds) modell hvis: $OR_{11} = OR_{10} \times OR_{01}$

$$\theta = OR_{11}/(OR_{10} \times OR_{01}) = \frac{ah/cf}{(bh/fg)(eh/df)} = adfg/bceh$$

* Både miljøeksponeringen og risikogenotypen er dikotome (0/1) variable. Størrelsene *a, b, c, d* osv. er de samme som i tabell 1.

at vi også kan estimere størrelsen på interaksjonen θ ved hjelp av logistisk regresjon når man har to dikotome studiefaktorer, noe som på en enkel måte også gir oss konfidensintervall for interaksjonsstørrelsen.

MODELLERING AV INTERAKSJONER MED LOGISTISK REGRESJON I CASE-KONTROLL DESIGN

Vi antar at den logistiske regresjonsmodellen er kjent (se evt. Kleinbaum et al. (8)). Her modelleres logaritmen til oddsen for sykdom som en lineær funksjon av forklaringsvariablene (gener, miljøfaktorer eller begge deler). Oddsen for sykdom er forholdet mellom sannsynligheten for sykdom (probability, Pr = risiko) og sannsynligheten for å ikke være syk (1 – sannsynligheten for å være syk). Hvis sannsynligheten for sykdom er liten er oddsen tilnærmet lik risikoen, og oddsratioen tilnærmet lik riskratioen (den relative risikoen). Det man utnytter i case-kontroll designet er at oddsratioen for sykdom forbundet med eksponering (en forklaringsvariabel) er lik oddsratioen for eksponering forbundet med sykdom. Nedenfor har jeg satt opp uttrykket for en logistisk regresjonsmodell med et konstantledd, to dikotome variable (x_1 og x_2) og et interaksjonsledd (multiplikasjonsledd, $x_1 \times x_2$) for interaksjonen mellom de to variablene:

$$ln\left[\frac{Pr(D=1|x_1,x_2)}{1-Pr(D=1|x_1,x_2)}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

Her kan x_1 være miljøeksponeringen mens x_2 kan være genotypen, begge for enkelthets skyld definert ved 1 (ja) eller 0 (nei). Her er β_1 og β_2 logistiske regresjonskoeffisienter for hovedeffektene av de to studiefaktorene. Disse koeffisientene eksponensiert (e^{β}) er da oddsratioen for de respektive faktorene. I denne situasjonen er regresjonskoeffisienten for interaksjonsleddet (β_3) eksponensiert (e^{β_3}) lik θ definert over. Dette kan enkelt utledes, se for eksempel Kleinbaum et al. (14).

Vi skal nå se på et eksempel med samlet effekt av allelisk variasjon i insulingen-regionen (INS VNTR) og HLA-DR risiko genotype (gen-gen interaksjon)

med hensyn på type 1 diabetes, hentet fra en case-kontroll studie beskrevet av Bain et al. (25). Resultatet vises i tabell 3 med en tradisjonell 2 × 2 × 2 tabell.

Vi ser at OR for de to strataene er omtrent like, og at forholdet mellom dem er 0,92, som er svært nær 1,0. Vi ser at også at konfidensintervallene omkring de to stratumspesifikke OR er ganske vide, og det følger derfor at det må være ganske stor usikkerhet rundt interaksjonsmålet θ. Vi skulle gjerne hatt et konfidensintervall for θ. Hvis vi tilpasser en logistisk regresjonsmodell med interaksjon mellom de to genene får vi ut estimater for β_3 og dens standardfeil (mange programpakker gir automatisk θ med konfidensintervall). Konfidensintervallet (95%) er da som vanlig gitt ved $e^{\beta_3 \pm 1.96 \times SE(\beta_3)}$. I eksemplet vårt får vi at $\beta_3 = -0{,}073$ med en standardfeil på 1,148, som gir oss $\theta = 0{,}93$ (95% konfidensintervall: 0,10 til 8,82). Siden en av cellene i tabell 3 bare har én observasjon, er det ikke overraskende at usikkerheten er stor (konfidensintervallet bredt). Strengt tatt bør man være forsiktig med å tolke konfidensintervallene fra både standard logistisk regresjon og tabellanalyse med Woolfs metode når noen av cellene har så få observasjoner. Selv om punktestimatene tyder på multiplikativ sammenheng i dette eksempelet, er dataene også konsistente med både sterk positiv og negativ interaksjon. Lav statistisk styrke er et generelt problem ved analyse av interaksjoner (26), og man må ofte basere seg på biologiske og andre faglige vurderinger.

CASE-ONLY DESIGN

Case-only design involverer, som navnet sier, kun cases og ingen kontroller. Case-only design kan ikke brukes til å estimere hovedeffekter av verken gener eller miljøfaktorer, men det kan brukes til å estimere graden av interaksjon mellom to faktorer i multiplikative risikomodeller hvis det er uavhengighet mellom prevalens av risikoallel og eksponering i befolkningen (blant kontrollene i en case-control studie).

Hvis man betrakter dataene fra en case-kontroll studie med hensyn til assosiasjonen mellom de to eksponeringene (gen og miljøfaktor eller de to genene) blant henholdsvis cases og kontroller kan dataene settes opp som i tabell 4, der *a, b, c, d, e, f, g* og *h* er de samme som i tabellene 1 og 2. Ved enkel algebra ser vi at man kan skrive $OR_{EG|D=1} = \theta \; x \; OR_{EG|D=0}$ (27), der $OR_{EG|D=1}$ er oddsratioen for assosiasjon mellom miljøfaktor og gen blant de syke (D=1), mens $OR_{EG|D=0}$ er oddsratioen for assosiasjon mellom miljøfaktor og gen blant kontrollene (D=0).

Tabell 4. 2 × 2 tabell for assosiasjon mellom to 'eksponeringer' (gen og miljøfaktor) stratifisert i henhold til kasus/kontroll status.

	Kasus (D=1)			Kontroller (D=0)		
	E=1	E=0		*E=1*	E=0	
G=1	*a*	*b*	G=1	*c*	*d*	
G=0	*e*	*f*	G=0	*g*	*h*	
	$OR_{EG\|D=1} = af/be$			$OR_{EG\|D=0} = ch/dg$		

$$\theta = OR_{EG|D=1} / OR_{EG|D=0} = (af/be)/(ch/dg) = adfg/bceh$$

$$\Downarrow$$

$$OR_{EG|D=1} = \theta \times OR_{EG|D=0}$$

EG indikerer at oddsratioen refererer til assosiasjonen mellom eksponering (E) og risikogenotype (G), OR_{EG}.

Man ser da at hvis prevalensen av risikoallelet og eksponeringen er uavhengig blant kontrollene, så blir $OR_{EG|D=0} = 1$, som medfører at $OR_{EG|D=1} = \theta$ og dermed kan man i prinsippet estimere θ uten å bruke informasjonen fra kontrollene. Hvis man har god grunn til å tro at det ikke er noen assosiasjon mellom det å bære et gen og eksponering i befolkningen, og uavhengigheten i tillegg bekreftes i en undersøkelse, så kan man altså effektivt estimere θ fra 'cases-only'. Man kan ofte forvente uavhengighet mellom et gen og en miljøfaktor i etnisk homogene befolkninger, men dette er slett ikke noen nødvendighet og det kan variere avhengig av miljøfaktor, gen og befolkning. Det er klart at miljøeksponeringer i seg selv normalt ikke påvirker hva slags gener en person bærer. Som regel er det også urimelig å tenke seg at ulike varianter av et gen kan påvirke eksponering, men man kan for eksempel tenke seg at eksponeringen alkoholinntak ikke ville være uavhengig av gener som påvirker nedbrytningen

Tabell 3. Eksempel med HLA, insulin gen (INS) og type 1 diabetes – 2 × 2 tabell stratifisert i henhold til HLA risikogenotype*.

	HLA risiko-genotype (G=1)			HLA 'normal' genotype (G=0)		
	INS +	INS –		INS +	INS –	
Kasus	120	25	Kasus	5	1	
Kontroller	134	80	Kontroller	45	28	
	$OR_{GD\|HLA=1}$= 2.87 (1.72, 4.78)			$OR_{GD\|HLA=0}$= 3.11 (0.45, 28.0)		

$\theta = 2.87/3.11 = 0.92$

* Eksempelet er tatt fra Bain et al. (25), tabell 4 og HLA-DR3, 4 og 3/4 er slått sammen til HLA risikogenotype ('HLA=1'), mens HLA-DRX/X er referansegenotype, der X er alleler annet enn DR3 eller 4 ('HLA=0').

av alkohol slik at de som bærer spesielle varianter lettere får ''bakrus'' når de drikker alkohol. Det kan følgelig også være avhengighet mellom en genetisk markør og en miljøeksponering hvis den markøren man måler er i koblingsulikevekt med spesielle varianter av gener som kan påvirke eksponeringen. Piegorsh og medarbeidere (27) har vist at estimering av interaksjonsstørrelsen θ gjøres mer presist enn i en standard case-kontroll analyse der man inkluderer kontrollene, gitt at denne forutsetningen om at $OR_{EG|D=0} = 1$ er oppfylt. Intuitivt kan man si at dette skyldes at man setter $OR_{EG|D=0}$ lik 1,0 i case-only analysen (etter å ha sjekket at dette er tilnærmet tilfelle), mens man i en standard case-kontroll analyse må estimere denne størrelsen og 'drar med seg' usikkerheten knyttet til $OR_{EG|D=0}$. Som for de fleste statistiske analyser ser vi at man kan øke den statistiske styrken ved å innføre flere antakelser. Case-only analyser kan med fordel brukes til en første undersøkelse av gen-gen interaksjoner. Hvis man har to gener (loci) som ligger på ulike kromosomer er det ofte god grunn til å tro at befolkningens prevalens av risikoalleler i ett locus er uavhengig av prevalensen av allelene i det andre locus. Selv om koblingsulikevekt strengt tatt ikke kan foreligge mellom alleler i loci på ulike kromosomer kan assosiasjon mellom alleler i slike loci oppstå hvis alleler i hvert av de to loci for eksempel er forbundet med seleksjon. Derfor bør man være forsiktig med å anta uavhengighet uten empirisk bekreftelse. Estimering av interaksjon med case-only analyser er svært sensitiv overfor avvik fra uavhengighetsantakelsen (28), og man må være klar over at en assosiasjon mellom de to risikofaktorene i befolkningen kan føre til enten overestimering eller underestimering av interaksjonsparameteren. I det siste tilfelle kan dette designet medføre tap av styrke som kanskje kan oppveie fordelen med økt presisjon i forhold til tradisjonell case-control analyse.

Å SLÅ SAMMEN KATEGORIER

Hittil har vi behandlet situasjoner for analyse av interaksjon mellom to ulike faktorer som begge har kun to nivåer (ja/nei). Ofte har man flere ulike gener (loci) og flere ulike miljøfaktorer som hver kan anta flere enn to nivåer, og som til sammen gir opphav til et stort antall mulige interaksjoner. Det første man bør gjøre er å begrense antall loci eller miljøfaktorer til de mest interessante med hensyn til interaksjoner. Dette bør være basert på faglige vurderinger om mekanismer og eventuelle tidligere studier (29,30).

Vi skal fortsatt begrense oss til situasjoner med to studiefaktorer, men hvor hver faktor nå kan anta flere enn to nivåer. Hvis vi har to gener som hver har to varianter, kan ni ulike kombinasjoner defineres (tabell 5).

Allerede på dette stadiet er situasjonen ganske kompleks. Analyse og tolkning av interaksjoner blir svært komplisert med økende antall nivåer for hver studiefaktor, og i de fleste situasjoner der en eller flere studiefaktorer har mer enn to nivåer finnes det ikke ett enkelt mål på interaksjon. I tillegg begynner problemet med utilstrekkelig utvalgsstørrelse å melde seg for alvor. For de fleste gener vil minst én av de ni kombinasjonene inneholde en svært liten andel av observasjonene, eller ikke inneholde noen observasjoner i det hele tatt. Da må man slå sammen grupper for å få til en meningsfylt analyse.

Avhengig av forekomsten av homozygote og heterozygote individer i befolkningen eller kunnskap om de biologiske egenskapene til genene kan man slå sammen grupper og definere 'eksponering' for risikoalleler som en ja/nei variabel på minst to ulike måter. På den ene side kan eksponering defineres som minst én kopi av 'risikoallelet' (Aa eller AA) versus ingen (aa), eller på den annen side som to varianter av allelet (AA) versus mindre enn to (Aa eller aa). Det er også mulig formelt å teste om ulike måter å slå sammen grupper på er rimelig ut fra statistiske betraktninger om modelltilpasning (27), men jeg skal ikke komme inn på dette her.

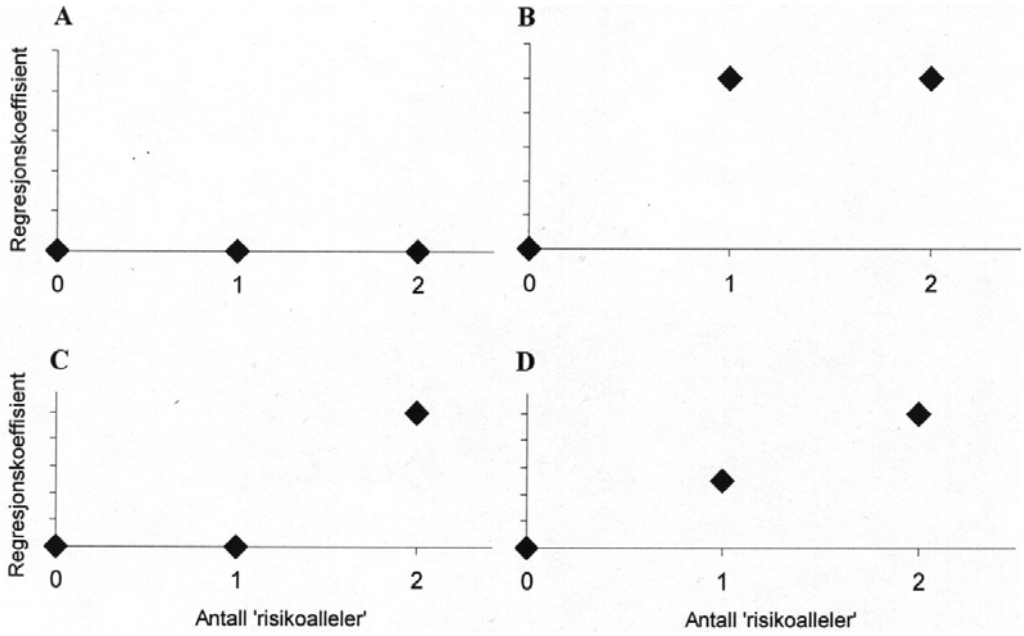
Tabell 5. Mulige kombinasjoner av 'eksponering' for to 'gener' (loci), hver med to varianter (alleler)*.

	Locus 1	Antall A alleler	Locus 2	Antall B alleler	Risikoalleler i locus 1 og 2
1	AA	2	BB	2	2,2
2	AA	2	Bb	1	2,1
3	AA	2	bb	0	2,0
4	Aa	1	BB	2	1,2
5	Aa	1	Bb	1	1,1
6	Aa	1	bb	0	1,0
7	aa	0	BB	2	0,2
8	aa	0	Bb	1	0,1
9	aa	0	bb	0	0,0

* To ulike alleler, A og a kan finnes i Locus 1, og to ulike alleler B og b kan finnes i Locus 2. 'Eksponering' for hvert locus er definert ved null, én eller to kopier av henholdsvis A og B. For hvert locus er ett allel arvet fra mor og ett arvet fra far.

DOSE-RESPONS SAMMENHENGER

Hvis man har en situasjon der det er ønskelig å ta hensyn til at gener eller eksponeringer kan anta mer enn to nivåer må man vurdere ('gen') dose-responssammenhengen. Selv om det ofte er umulig å klart definere en klassisk Mendelsk arvegang (f.eks. autosomal recessiv eller autosomal dominant) for komplekse sykdommer, er man ofte interessert i å estimere hvordan sykdomsrisikoen øker med 'gendose' (alleldose). Figur 1 viser fire typiske dose-respons mønstre, der den logistiske regresjonskoeffisienten er plottet mot antall 'risikoalleler'. I praksis avviker observasjonene som regel noe fra disse mønstrene (kalles i kvantitativ genetikk for dominansavvik). Det er oftest en fordel å starte med en modell der man gjør færrest mulig antakelser om disse mønstrene.



Figur 1. Fire typiske gen-dose respons mønstre. Regresjonskoeffisienten fra logistisk regresjon er plottet mot antall risikoalleler. A: Ingen effekt, B: 'Dominant' effekt, C: 'Recessiv effekt', D: 'Gen-dose' effekt.

Det betyr at man først estimerer effekten av en kategorisk variabel (de fleste programmer lager da såkalte "dummy-variable"). Man må definere en referansekategori som typisk vil være "null risikoalleler". Da vil den første logistiske regresjonskoeffisienten estimere effekten av ett allel sammenlignet med null risikoalleler, mens den andre regresjonskoeffisienten estimerer effekten av to risikoalleler sammenlignet med null. Regresjonskoeffisienten for to alleler vs. ett estimeres ikke direkte. Den er definert ved differansen mellom den andre og den første regresjonskoeffisienten. Den enkleste måten å estimere den siste størrelsen på med konfidensintervall er å velge en annen referansekategori og kjøre analysen på nytt.

Hvis man finner et tilnærmet lineært forhold mellom antall alleler og den logistiske regresjonskoeffisienten som i figur 1D er det rimelig å estimere effekten av genet med antall alleler som en kontinuerlig variabel i logistisk regresjon (tre nivåer; 0, 1 eller 2). Da behøver man kun å estimere én regresjonskoeffisient for effekten av genet, istedenfor to. Det betyr at sykdomsrisikoen (-oddsen) øker like mye fra null til én som fra én til null. Regresjonskoeffisienten for to alleler sammenlignet med null er gitt ved $e^{2\beta}$, der β er regresjonskoeffisienten i modellen der antall alleler er brukt som kontinuerlig variabel. Siden logistiske regresjonsmodeller er multiplikative risiko- (odds-) modeller, kan man si at det er multiplikativ alleldose-responssammenheng.

Hvis observasjonene viser at effekten av to alleler er så mye større enn effekten av ett allel at det er urimelig å anta at sammenhengen mellom antall alleler og regresjonskoeffisienten er lineær, kan man si at man har en 'overmultiplikativ' alleldose-responssammenheng (dominansavvik i lineære modeller). Enkelte snakker da om 'intra-lokus' interaksjon. Selv om hvert allel i et locus nedarves uavhengig av hverandre, kan altså effekten på sykdomsrisiko (f.eks. relativ risiko) forbundet med en økning fra null til én risikoalleler være annerledes enn økningen fra én til to risikoalleler i en gitt modell. På lignende vis som for interaksjon mellom to ulike loci (gen-gen interaksjon, eller inter-locus interaksjon) som diskutert over, kan man også definere 'fravær av interlokusinteraksjon' i henhold til både additive og multiplikative risikomodeller. Når man bruker logistisk regresjon (eller tabellanalyse med oddsratioer eller relativ risiko), er man i en multiplikativ risiko (odds) modell, og man må være klar over at begrep som additiv og multiplikativ gen-dose effekt er avhengige av hva slags skala eller modell man bruker for effekt. Dette er beslektet med skala-avhengigheten, eller modellavhengigheten til gen-gen eller gen-miljø interaksjon.

Ofte vil observasjonene avvike noe fra de typiske mønstrene i figur 1, og man bør plotte regresjonskoeffisientene med 95% konfidensintervall (f.eks. ved å bruke Excel) for å se hvor stor usikkerheten i dose-responsmønsteret er. Det er også mulig å formelt teste om én dose-respons modell er signifikant bedre enn en annen, men jeg skal ikke komme inn på dette her. Som regel må biologiske og andre faglige vurderinger tas hensyn til sammen med statistisk usikkerhet.

Å TESTE FLERE HYPOTESER SAMTIDIG

Når man vil teste mange hovedeffekter og interaksjoner i samme studie kan det lett oppstå problemer med tolkning av statistisk signifikans. Hvis man har å gjøre

med et gen hvor en hovedeffekt ennå ikke er etablert kan man i samme datasett være interessert i å teste for både hovedeffekt (assosiasjon) og interaksjon. I en slik situasjon kan det være fornuftig å gjøre en slags 'klareringstest', der man tester flere hypoteser om både assosiasjon og interaksjon samtidig (31,32). Dette er analogt til variansanalyse-situasjonen, der man først tester om det er minst én effekt, og går videre med å finne de spesifikke effektene hvis klareringstesten gir signifikant resultat. Utskriften fra logistisk regresjon kan brukes til likelihood ratio tester for flere hypoteser samtidig. For enkelte programpakker vil det da være nødvendig med noe håndregning, og jeg vil ikke gå inn på detaljene her. Interesserte lesere kan for eksempel se i boka til Kleinbaum og medarbeidere (8). Man kan for eksempel teste interaksjoner og hovedeffekter av to gener samtidig ved å sammenligne modellen med parametre for alle disse effektene (pluss et konstantledd) med en modell uten hovedeffekter eller interaksjoner (bare konstantledd, alle regresjonsparametrene er lik 0). Hvis en slik klareringstest er signifikant kan man systematisk estimere hovedeffekter og eventuelle interaksjoner. Selv om mange studier av gen-miljøinteraksjoner har involvert både gener og miljøfaktorer som allerede er etablerte eller tidligere studerte som risikofaktorer for sykdom, kan det godt tenkes at interaksjonseffekter kan påvises uten at det kan påvises hovedeffekt (assosiasjon) for minst én av enkeltfaktorene.

DISKUSJON

De fleste av prinsippene som er omtalt her er hentet fra grunnleggende epidemiologi og statistikk. Jeg har i denne artikkelen begrenset meg til case-kontroll-designet. De samme prinsippene som er omtalt her gjelder også for kohortdesign. Kohortdesign har den fordelen at miljøeksponeringer ofte kan måles bedre enn i case-kontroll studier. Med kohortdesign kan man også estimere interaksjon i henhold til additive modeller, mens man i case-kontroll studier i hovedsak er begrenset til multiplikative risiko (odds) modeller. Kort fortalt skyldes dette at man kun kan estimere relativ risiko (oddsratio) i case-kontroll studier, og ikke absolutt risiko eller risikodifferens. Rothman har foreslått noen knep for å vurdere interaksjon i henhold til additiv risikomodell også i case-kontroll studier (13,18,33). Disse metodene har sine begrensninger, som jeg ikke skal komme inn på her (Anders Skrondal. Measures of interaction in case-control studies with covariates: a cautionary note. Innsendt manuskript). Case-kontroll designet er langt billigere og enklere å gjennomføre. De fleste gener endrer seg ikke med tiden, og kan derfor måles like godt etter at sykdom er oppstått, som i case-kontroll studier. Når interaksjon i en case-kontroll studie betraktes i henhold til et case-only oppsett (som i tabell 4) ser man at estimering av interaksjonsparameteren θ ikke påvirkes av eventuell systematisk forskjell i eksponeringsmåling mellom kasus og kontroller (32). Man bør allikevel ikke forledes til å tro at case-kontroll studier er enkle. Nyanser i valg av kontroller, matching, definering av eksponering og analyse er ofte langt mer kompliserte enn hva mange tror (34). Andre design og analysemetoder, som koblingsstudier og analyser av pasient-foreldre triader kan også brukes til å studere interaksjoner basert på beslektede prinsipper. Alle design har sine ulemper og fordeler (6,7,35). Clayton & KcKeigue har nylig hevdet at gen-gen og gen-miljøinteraksjoner ved komplekse sykdommer har begrenset vitenskapelig verdi, men at hvis man først skal undersøke slike interaksjoner med epidemiologiske design, så er case-kontroll designet å foretrekke (32). En av hovedinnvendingene Clayton & McKeigue har mot studier av interaksjon er at spørsmålet om interaksjon ofte er avhengig av hva slags modell man velger for å definere fravær av interaksjon. Det må da legges til at i situasjoner der minst én av enkeltfaktorene ikke har effekt alene, så er spørsmålet om interaksjon ikke avhengig av modellvalg. Det er mulig å se på studier av interaksjon som forsøk på å beskrive dose-respons effekten av ulike studiefaktorer. Slik kan det ha betydning for risikoprediksjon, design av studier og tolkning av studier av miljøeksponeringer hvor man begrenser seg til personer med risikogenotyper (eller sykdom i familien). Jeg tror vi i de neste årene vil se et økende antall studier av gen-gen og gen-miljøinteraksjoner. Det er derfor viktig å være klar over tolkningsmuligheter og -begrensninger ved slike studier.

I studier av molekylærgenetiske markører er det svært viktig med god kommunikasjon mellom personer med kompetanse i epidemiologiske metoder og personer med kompetanse i molekylærbiologiske metoder for å oppnå en meningsfylt analyse. I denne artikkelen har jeg ignorert en rekke komplekse biologiske forhold som kan ha betydning for planlegging og tolkning av studier. Ett eksempel er muligheten for at alleler kan ha ulik effekt avhengig av om de arves fra mor eller far. For å vurdere dette er det nødvendig å analysere DNA fra foreldrene til 'indekspersonene'. Videre kan eksterne faktorers regulering av genenes uttrykking (ekspresjon) være viktigere for sykdomsrisiko enn allelisk variasjon. I alle studier av geners assosiasjon med sykdom bør man forsikre seg om at allelene er i såkalt Hardy-Weinberg likevekt i befolkningen, ellers blir resultatene vanskelig å tolke (1). Jeg har ikke kommet inn på dette i denne artikkelen, men anbefaler at man setter seg inn i hva dette innebærer hvis man skal gjøre analyser som beskrevet her på virkelige datasett. Et eksempel på en omfattende lærebok i human genetikk for interesserte lesere er Vogel & Motulsky (36). En viktig begrensing som ofte diskuteres i forbindelse med case-kontroll studier av genetisk assosiasjon er såkalt befolkningsstratifisering (35,37). Dette betyr at dersom befolkningen for eksempel består av ulike etniske grupper som har både ulik prevalens av risikoallelet og sykdomsinsidens, så kan ikke-kausale assosiasjoner oppstå. Dette er en av

hovedbegrunnelsene for bruk av familiebaserte studier. Allikevel mener enkelte at frykten for feil forårsaket av slik befolkningsstratifisering er overdrevet (38). Fenomenet kan betraktes som tradisjonell confounding som kan kontrolleres i design eller analyse hvis man kan måle den relevante stratifiseringen (for eksempel etnisk gruppe eller ved kjente, uavhengige genetiske markører).

Til slutt vil jeg si at en av de viktigste praktiske begrensningene ved studier av interaksjon er utilstrekkelig statistisk styrke (26). Av den grunn er det viktig at man gjør nøye vurderinger av design, analysemetoder og styrkeberegninger før man setter i gang en studie med primær målsetning å studere interaksjoner. Ulike varianter av matching (39) og såkalt countermatching (40) har vært foreslått for å øke den statistiske styrken i spesielle situasjoner, men hver studie krever sine spesielle hensyn.

ACKNOWLEDGMENT

Jeg vil takke Anders Skrondal, Kjersti Skjold Rønningen, Per Magnus og en anonym referee for nyttige kommentarer på tidligere utkast til denne artikkelen.

REFERANSER

1. Khoury MJ, Beaty TH, Cohen BH. *Fundamentals of genetic epidemiology*. New York: Oxford University Press, 1993.
2. Elston RC. Introduction and overview. Statistical methods in genetic epidemiology. *Stat Methods Med Res* 2000; **9**: 527-41.
3. Ellsworth DL, Manolio TA. The emerging importance of genetics in epidemiologic research. I. Basic concepts in human genetics and laboratory technology. *Ann Epidemiol* 1999; **9**: 1-16.
4. Ellsworth DL, Manolio TA. The emerging importance of genetics in epidemiologic research. II. Issues in study design and gene mapping. *Ann Epidemiol* 1999; **9**: 75-90.
5. Ellsworth DL, Manolio TA. The emerging importance of genetics in epidemiologic research. III. Bioinformatics and statistical genetic methods. *Ann Epidemiol* 1999; **9**: 207-24.
6. Yang Q, Khoury MJ. Evolving methods in genetic epidemiology. III. Gene-environment interaction in epidemiologic research. *Epidemiol Rev* 1997; **19**: 33-43.
7. Andrieu N, Goldstein AM. Epidemiologic and genetic approaches in the study of gene-environment interaction: an overveiw of available methods. *Epidemiol Rev* 1998; **20**: 137-47.
8. Kleinbaum DG, Kupper LL, Muller KE, Nizam A. *Applied regression analysis and multivariable methods, 3rd edn*. Pacific Grove: Duxbury Press, 1998.
9. Haverkos HW. Could the aetiology of IDDM be multifactorial? *Diabetologia* 1997; **40**: 1235-40.
10. Lucchinetti C, Rodriguez M, Weinshenker BG. Multiple sclerosis. *N Engl J Med* 2000; **343**: 938-52.
11. Cookson W. The alliance of genes and environment in asthma and allergy. *Nature* 1999; **402** (Suppl):B5-B11.
12. Ross R. Atheriosclerosis – an inflammatory disease. *N Engl J Med* 1999; **340**: 115-26.
13. Rothman KJ. Synergy and antagonism in cause-effect relationships. *Am J Epidemiol* 1974; **99**: 385-88.
14. Kleinbaum DG, Kupper L, Morgenstern H. Interaction, effect modification, and synergism. In: Kleinbaum DG, Kupper L, Morgenstern H, editors. *Epidemiologic research: principles and quantitative methods*. New York: Wiley & Sons, 1982: 403-18.
15. Greenland S, Rothman KJ. Concepts of interaction. In: Rothman KJ, Greenland S, editors. *Modern epidemiology*. Philadelphia: Lippincott-Rave, 1998: 329-42.
16. Risch NJ. Searching for genetic determinants in the new millenium. *Nature* 2000; **405**: 847-56.
17. Weinberg CR. Applicability of the simple independent action model to epidemiologic studies involving two factors and a dichotomous outcome. *Am J Epidemiol* 1986; **123**: 162-73.
18. Rothman KJ. *Modern epidemiology*. Boston: Little, Brown & Co., 1986.
19. Siemiatycki J, Thomas DC. Biological models and statistical interactions: an example from multistage carcinogenesis. *Int J Epidemiol* 1981; **10**: 383-7.
20. Thompson WD. Effect modification and the limits of biological inference from epidemiologic data. *J Clin Epidemiol* 1991; **44**: 221-32.
21. Cordell HJ, Todd JA, Hill NJ, Lord CJ, Lyons PA, Peterson LB, et al. Statistical modelling of interlocus interactions in a complex disease. Rejection of the multiplicative model of epistasis in type 1 diabetes. *Genetics* 2001; **158**: 357-67.
22. Breslow NE, Day NE. *Statistical methods in cancer research, Vol. 1, the analysis of case-control studies*. IARC Scientific Publications No. 32. Lyon: International Agency for Research on Cancer, 1980.
23. Kleinbaum DG, Kupper L, Morgenstern H. *Epidemiologic research: principles and quantitative methods*. New York: John Wiley & Sons, 1982.

24. Botto LD, Khoury MJ. Commentary: facing the challenge of gene-environment interaction: the two-by-four table and beyond. *Am J Epidemiol* 2001; **153**: 1016-20.
25. Bain SC, Prins JB, Hearne CM, Rodrigues NR, Rowe BR, Pritchard LE, et al. Insulin gene region-encoded susceptibility to type 1 diabetes is not restricted to HLA-DR4-positive individuals. *Nat Genet* 1992; **2**: 212-5.
26. García-Closas M, Lubin HJ. Power and sample size calculations in case-control studies of gene-environment interactions: comments on different approaches. *Am J Epidemiol* 1999; **149**: 689-92.
27. Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Stat Med* 1994; **13**: 153-62.
28. Albert PS, Ratnasinghe D, Tangrea J, Wacholder S. Limitations of the case-only design for identifying gene-environment interactions. *Am J Epidemiol* 2001; **154**: 687-93.
29. Greenland S. Modelling and variable selection in epidemiologic analysis. *Am J Public Health* 1989; **79**:340-9.
30. Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996; **15**: 361-87.
31. Longmate JA. Complexity and power in case-control association studies. *Am J Hum Genet* 2001; **68**: 1229-37.
32. Clayton D, McKeigue PM. Epidemiological methods for studying genes and environmental factors in complex diseases. *Lancet* 2001; **358**: 1356-60.
33. Rothman KJ. The estimation of synergy or antagonism. *Am J Epidemiol* 1976; **103**: 506-11.
34. Rothman KJ, Greenland S. *Modern epidemiology, 2nd edn*. Philadelphia: Lippincott-Raven, 1998.
35. Weinberg CR, Umbach DM. Choosing a retrospective design to assess joint genetic and environmental contributions to risk. *Am J Epidemiol* 2000; **152**: 197-203.
36. Vogel F, Motulsky AG. *Human genetics: problems and approaches, 3rd edn*. Berlin: Springer, 1997.
37. Knowler WC, Williams RC, Pettitt DJ, Steinberg AG. Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. *Am J Hum Genet* 1988; **43**: 520-6.
38. Wacholder S, Rothman N, Caporaso N. Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. *J Natl Cancer Inst* 2000; **92**: 1151-8.
39. Sturmer T, Brenner H. Potential gain in efficiency and power to detect gene-environment interactions by matching in case-control studies. *Genet Epidemiol* 2000; **18**: 63-80.
40. Andrieu N, Goldstein A, Thomas D, Langholz B. Counter-matching in studies of gene-environment interaction: efficiency and feasibility. *Am J Epidemiol* 2001; **153**: 265-74.