Analyse av pasient-foreldre triader;
en praktisk gjennomgang

Rolv T. Lie[1,2] og Astanand Jugessur[1,3]

1. Seksjon for medisinsk statistikk, Institutt for samfunnsmedisinske fag, Universitetet i Bergen
2. Medisinsk fødselsregister, Universitetet i Bergen
3. Senter for medisinsk genetikk og molekylærmedisin, Haukeland sykehus

Korresponderende forfatter: Rolv T. Lie, rolv.lie@smis.uib.no

SAMMENDRAG

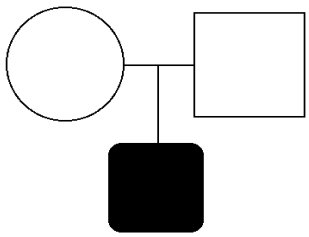
Stadig flere epidemiologiske studier har innsamling av genetisk materiale som en viktig komponent. Før studien settes i gang gjøres det vanligvis nøye vurderinger av designalternativer og av forskjellige praktiske løsninger. Målet vil alltid være å få størst mulig materiale med minst mulig feil med de ressursene man har tilgjengelig. Ofte er det kontrollmaterialet som bidrar med størst usikkerhet og høyest kostnad. I denne artikkelen gjennomgår vi noen av de mulighetene som ligger i å samle inn data fra pasienter og deres biologiske foreldre (pasient-triader) uten å samle inn et tradisjonelt kontrollmateriale. Med et slikt materiale kan man studere effekter av alleler og samspill mellom alleler og miljøeksponering med mindre risiko for feil enn ved andre epidemiologiske designalternativer. I forhold til pasient-kontroll og kohortstudier er hovedeffekter av eksponering det eneste man ikke kan studere. De statistiske metodene kan gjennomføres med vanlig statistisk programvare. Vi viser en rekke eksempler på analyser ved hjelp av programpakken STATA. Som eksempel har vi brukt et foreløpig sett av pasient-triader fra SAM-prosjektet, som er en nasjonal studie av leppe- og ganespalte. Analysene er kun ment som en illustrasjon av de aktuelle statistiske metodene.

INNLEDNING

De fleste epidemiologer har etter hvert vent seg til å tenke på gener som en form for eksponering som kan studeres på vanlig måte i pasient-kontrollstudier eller kohortstudier. Når det startes en større epidemiologisk feltstudie med direkte innsamling av spørreskjemaopplysninger eller andre opplysninger fra deltagerne vurderes det nøye om det også skal samles inn biologisk materiale som kan benyttes til genotyping. Faglige hensyn vil ofte tale for, mens praktiske og etiske hensyn kan tale imot. Det naturlige i en tradisjonell epidemiologisk tilnærming ville være å samle inn biologisk materiale fra pasientene og fra kontrollene i en pasient-kontroll studie eller fra alle deltagere i en kohortstudie.

Det mest overraskende ved de nye metodene som presenteres her er at det er mulig å studere effekter av gener og av interaksjon mellom gener og eksponering uten egentlig å ha noen kontrollgruppe. Ved å konsentrere innsatsen om pasientgruppen og samle inn eksponeringsinformasjon samt biologisk materiale fra pasienten og fra begge foreldrene (pasient-triade, fig. 1) kan man gjøre en studie som har færre feilkilder samt gir mulighet til å undersøke mer kompliserte genetiske effekter, for eksempel genomisk imprinting (1). Antagelig er det eneste som ikke kan estimeres fra en slik studie hovedeffekter av miljøeksponering.

Alle epidemiologer som har deltatt i større feltstudier vet at hovedutfordringen ofte består i å få samlet inn gode data fra en representativ kontrollgruppe. Når man skal gjøre en assosiasjonsstudie med gener blir ikke disse problemene mindre. Det faktum at man kan gjøre en bedre studie kun med utgangspunkt i en pasientgruppe er oppsiktsvekkende. Denne gruppen er ofte motivert og kontakt kan ofte formidles på en naturlig måte gjennom en behandlende avdeling. En målrettet rekruttering av pasienter og deres foreldre er ofte svært rimelig i forhold til å sette opp et stort apparat for å etablere en kohort eller for å samle inn data på en representativ kontrollgruppe.



Figur 1. Mor, far og barn triade. Vanligvis vil barnet i triaden være pasient.

En viktig begrensning for disse metodene er at det må være mulig å få tak i biologisk materiale både fra pasienten og fra de biologiske foreldrene til pasienten. Dette er opplagt vanskelig når pasientene selv er gamle. Relevansen av de spesielle fordelene ved triadedesignet som estimering av direkte effekter av mors alleler og imprinting er også størst for sykdom hos

barn og nyfødte. Designet er faktisk spesielt godt egnet til å studere svangerskapsutfall ved at man kan studere de separate genetiske bidrag fra de tre partene i et svangerskap: mor, far og barn. Forskere som er opptatt av at svangerskapet har avgjørende betydning for helse også i voksen alder, for eksempel ifølge den såkalte Barkerhypotesen, vil allikevel kunne argumentere for å studere bidrag av mors alleler og imprinting også for sykdom hos voksne. Selv om det ikke er sikkert at pasienten lenger er et barn, kommer vi til å referere til de tre personene i triaden som mor, far og barn.

I denne artikkelen vil vi gi en del praktisk veiledning i hvordan man kan analysere data fra pasient-triader. Hovedfokus vil være på å vise hvordan man kan estimere effekter av alleler fra pasient-triader. Ved hjelp av vanlig statistisk programvare, her Poisson-regresjon i STATA (2), vil vi vise at det går an å estimere effekter av alleler hos barnet og alleler hos mor separat (3,4). Vi viser også hvordan man kan studere interaksjon mellom alleler og eksponering (5). I tillegg demonstrerer vi hvordan det kan gjøres en test for imprinting ved hjelp av logistisk regresjon (6). Vi gir noen råd om innsamling og behandling av biologisk materiale i store epidemiologiske studier basert på egne erfaringer og planer. Praktiske eksempler vil være hentet fra prosjektet svangerskap, arv og miljø (SAM-prosjektet) som egentlig er en pasient-kontrollstudie av leppe- og ganespalte men som inneholder pasient-triade data. Eksemplene er basert på foreløpige data fra et lite utvalg av materialet. Vi presenterer kun data for leppe- og ganespalte samlet, mens det i en endelig analyse vil være nødvendig å skille ut pasienter med isolert ganespalte som en separat gruppe.

Vi beskriver analyser for en situasjon hvor det aktuelle genet kun har to alleler. Dersom genet har flere alleler, vil man i praksis måtte analysere et og et allel i forhold til de andre. Utviklingen går i retning av at en meningsfull genetisk analyse også bør inkorporere flere gener. Det lar seg antagelig relativt lett gjøre å generalisere disse metodene til å se på flere gener og samspill mellom genene. Det er imidlertid ikke opplagt at regresjonsmetodene som presenteres her er ideelle for analyse av et stort antall gener samtidig. Her ligger det åpenbart et behov for videreutvikling av de statistiske metodene.

INNSAMLING AV GENETISK MATERIALE

Den største praktiske utfordringen i en pasient-triade studie er innsamling av biologisk materiale som kan brukes til ekstraksjon av DNA. Det er klart at innsamling av blod oftest vil være å foretrekke både på grunn av mengden DNA og på grunn av kvaliteten. For pasienter som er under behandling vil det ofte finnes både anledning og motivasjon til at det kan avgis en blodprøve. Foreldre som følger et barn til behandling på et sykehus vil også ofte være motivert for å avgi blodprøve. I SAM-prosjektet har deltagelsen i pasientgruppen ligget på over 90%. Blant de som har samtykket til å delta har over 95% av fedrene og mødrene begge avgitt blodprøve. De fleste gjør dette på sykehuset i forbindelse med at barnet får behandling, men en del fedre har fått tatt blodprøven hos sin egen lege. Denne høye oppslutningen kan ha mange grunner, men gjenspeiler nok hovedsakelig at foreldrene ønsker å engasjere seg for at det skal kunne fremkomme ny kunnskap. Vi tror også at det har betydning at alle data, inklusive de biologiske prøvene fra prosjektet anonymiseres før de analyseres.

For enkelte studier vil det være praktisk vanskelig å samle inn blodprøver. I SAM-prosjektet samles det også inn munnhuleprøver fra deltagere i en kontrollgruppe og fra søsken av pasienten. Dette gjøres ved at det sendes et sett med Q-tips lignende bomullspinner sammen med en liten plastbeholder med isopropanol. Bomullspinnene strykes mot innsiden av kinnet et par ganger før de puttes opp i beholderen som lukkes og sendes pr. post tilbake til prosjektet. Denne metoden har fungert godt i praksis, men det er fortsatt usikkert hvor mye DNA som kan ekstraheres fra et slikt sett med bomullspinner. Et anslag går på at det kan være nok til analyse av et par hundre gener.

En annen metode ser imidlertid ut til å bli en standardmetode for storskala innsamling av DNA for epidemiologiske studier (7). Den såkalte ”mouthwash-metoden” består i at man skyller munnen med et spesiallaget ”munnvann” som så spyttes tilbake i en beholder. Denne metoden er vist å være bedre enn en munnbørste-metode (8). En del praktiske råd for hvordan man skal skaffe seg best mulig resultat av innsamlingen er nylig publisert av Feigelson et al. (9).

DNA-mengden vil kunne være en begrensende faktor ved denne typen innsamling av DNA. Det er viktig å huske at i studier som baserer seg på anonymisering vil det være svært vanskelig å kunne gå tilbake til deltagere for å be om mer DNA. Et nytt interessant alternativ er å formere opp hele den begrensede DNA-mengden man har fått innhentet ved hjelp av såkalt ”whole genome amplification”. I teorien vil man da kunne ha nesten ubegrensede mengder DNA tilgjengelig selv om man for eksempel kun har samlet inn spyttprøver. Det finnes en bekymring for at ”whole genome amplification” kan ødelegge mulighetene for å gjøre enkelte typer genetiske analyser, men for mye av det som er aktuelt i dag ser det ut til å fungere (10).

PROBLEMET MED BEFOLKNINGSSTRATIFISERING

Det grunnleggende problemet med pasient-kontrollstudier av genetiske effekter er risikoen for at effektene fordreies ved at man kan ha såkalt befolkningsstratifisering. Dette kan for eksempel oppstå når befolkningen som studeres er heterogen ved at en del av befolkningen har høy risiko for sykdom av andre grunner samtidig som de har høy forekomst av allelet som studeres. I en pasient-kontrollstudie vil dette gi positiv bias (effektfordreining) og en falsk tendens til

positiv assosiasjon mellom sykdom og allel ved at allelet er overrepresentert blant pasientene. Dette har vært brukt som en god grunn til å gjøre studier i relativt homogene befolkninger som den norske. Problemet kan også oppfattes som en type confounding (effektforveksling) etnisitet som det delvis kan justeres for dersom man har informasjon om hvilke etniske grupper deltagerne kommer fra. Nylig ble det påpekt at problemet med befolkningsstratifisering kanskje ikke er så stort, og at gode pasient-kontrollstudier vil ha en viktig plass også i fremtiden (11).

TDT-TESTEN

På bakgrunn av problemene med befolkningsstratifisering beskrev Falk og Rubinstein allerede på 80-tallet muligheten av å bruke foreldrenes genotyper i en triade-design (12). De arbeidet med diabetes og HLA-typer, og ønsket å unngå problemene med å finne representative kontroller. De oppdaget at de haplotypene (eller genotypene) som ikke var blitt overført fra foreldre til barnet kunne egne seg som en kontroll i forhold til de haplotypene som var overført. Tilsynelatende uavhengig av dette presenterte også en annen gruppe HLA-forskere en tilsvarende analysestrategi i 1991 (13).

Den egentlige og enkle formuleringen av TDT-testen kom med Spielman et al. i 1993 (14). Her beskrives triadene som en matchet design og analysene gjøres ved en såkalt McNemar-test. Denne testen er beskrevet i de fleste lærebøker i biostatistikk.

I tabell 1 sees 4 forskjellige par av foreldre hvor det alltid vil være mulig å bestemme for den ene av foreldrene hvilket allel som ikke ble overført til barnet. For linje 1 er det klart at dersom barnet er *AA*, vil allelet *a* fra far ikke være overført og vil være en matsjet kontroll for *A* som ble overført fra far. Fra mor vil det alltid overføres en *A*. Tilsvarende for de andre kombinasjonene. Det er klart at i et matsjet design er det bare de disjunkte parene som er informative og i McNemar-testen er det bare disse observasjonene som inngår i beregningen. Beregningene består derfor i å telle opp antall ganger vi vet at *A* og ikke *a* overføres, og antall ganger *a* og ikke *A* overføres. Denne opptellingen kan man gjøre for kombinasjoner av foreldre hvor den ene er homozygot (*aa* eller *AA*) og den andre av foreldrene er heterozygot (*Aa*). Familier hvor begge foreldrene er homozygote eller begger er heterozygote er ikke informative.

Tabell 1. Kombinasjonene av alleler hos mor, far og barn som benyttes i den vanlige TDT-testen.

Mor	Far	Barn
AA	**Aa**	A**A** eller A**a**
Aa	AA	**A**A eller **a**A
aa	**Aa**	a**A** eller a**a**
Aa	aa	**A**a eller **a**a

I tabell 2 sees fordelingen av triadetyper for TGFα-genet etter genotyping av 122 triader. *A* betegner her det sjeldne eller muterte allelet. Til sammen 43 er triader informative for TDT-testen. Legg merke til at når man sammenligner linje 5 mot 6 og 7 mot 8 i tabellen vil man få en test på om det første overførte allelet til barnet har effekt (tegn på dominant effekt). Når man sammenligner linje 1 med 2 og linje 3 med 4 tester man om det andre allelet A som overføres til barnet har effekt utover det første (gen-dose-effekt). I den vanlige TDT-testen legges begge disse effektene inn i beregningen. Derfor er også TDT-testen relativt god når det er en gen-dose effekt, men ikke så god dersom det er andre typer effekter.

TDT-testen beregnes som en McNemar test:

$$\chi_1^2 = (17\text{-}26)^2/(17 + 26) = 1{,}88,\ p=0{,}17$$

Ved hjelp av et program for eksakte statistiske metoder kan man få en såkalt eksakt p-verdi for TDT-testen. Med programmet Statxact (15) blir p-verdien 0,22. Det er her altså ikke noen tendens til at allelet *A* overføres oftere til pasientene enn allelet *a*.

Tabell 2. Fordeling av informative triade-typer for *TGFα*-genet, 122 leppe- eller ganespalte pasienter.

Mor	Far	Barn	Antall	O	IO	Overføring fra –
AA	Aa	AA	0	A	a	Far
AA	Aa	Aa	0	a	A	Far
Aa	AA	AA	1	A	a	Mor
Aa	AA	Aa	1	a	A	Mor
Aa	aa	Aa	9	A	a	Mor
Aa	aa	aa	9	a	A	Mor
aa	Aa	Aa	16	A	a	Far
aa	Aa	aa	7	a	A	Far

O: overført IO: ikke overført

Tabell 3. Tabell for beregning av TDT-testen for *TGFα*.

		"Ikke overført" A	"Ikke overført" a
"Overført"	A	–	26
	a	17	–

Prinsippet med å bruke allelene som matsjede kontroller som i TDT-testen er blitt benyttet til å utvikle metoder for analyse av interaksjon med miljøeksponering. Det finnes et slikt eksempel på bruk av betinget logistisk regresjon (6).

15 TRIADETYPER OG 6 KRYSNINGSTYPER

I stedet for å prøve å ta ut spesielt informative triader som i TDT-testen, kan det være en god ide å se på hele fordelingen av alle 15 mulige triadetyper (3). Dersom man ser på allelfordelingen hos foreldrene kan det de-

fineres 6 krysningstyper. Disse er vist i tabell 4. I krysningstype 1 er begge foreldrene *AA*. Krysningstype 2 er definert ved at den ene av foredrene er *AA* og den andre er *aA* og så videre. Krysningstypene er nesten definert av hvor mange *A*-alleler foreldrene har. Unntaket er de forskjellige typene 3 og 4 hvor foredrene til sammen har to *A*-alleler.

Tabell 4. Oversikt over de 15 forskjellige triadetypene.

MFB[1]	KT[2]	H-W[3] fordeling	*Generell* fordeling
2 2 2	1	p^4	μ_1
2 1 2	2	$p^3(1-p)$	μ_2
2 1 1	2	$p^3(1-p)$	μ_2
1 2 2	2	$p^3(1-p)$	μ_2
1 2 1	2	$p^3(1-p)$	μ_2
2 0 1	3	$p^2(1-p)^2$	μ_3
0 2 1	3	$p^2(1-p)^2$	μ_3
1 1 2	4	$p^2(1-p)^2$	μ_4
1 1 1	4	$2p^2(1-p)^2$	$2\mu_4$
1 1 0	4	$p^2(1-p)^2$	μ_4
1 0 1	5	$p(1-p)^3$	μ_5
1 0 0	5	$p(1-p)^3$	μ_5
0 1 1	5	$p(1-p)^3$	μ_5
0 1 0	5	$p(1-p)^3$	μ_5
0 0 0	6	$(1-p)^4$	μ_6

[1] Kolonnen angir antall *A*-alleler hos henholdsvis mor, far og barn
[2] Krysningstyper av foreldre
[3] Forventet fordeling ved Hardy-Weinberg likevekt

Dersom man kan anta Hardy-Weinberg likevekt for de aktuelle allelene *A* og *a*, og at valg av partner er uavhengig av hvor mange *A*-alleler de har, kan fordelingen av krysningstyper settes opp ved hjelp av allelfrekvensen *p* for allelet *A* (også i det videre er *p* hyppigheten av allelet *A* og *1-p* hyppigheten av allelet *a*). Da blir frekvensen av krysningstypene 3 og 4 sammenfallende. Legg merke til at dette er fordelingen av krysningstypene i befolkningen, ikke i pasientgruppen. Dersom overføringen av alleler til barnet følger Mendelske lover, vil også hver triadetype innenfor hver krysningstype være like hyppige.

Dersom det ikke kan antaes at allelet er i Hardy-Weinberg likevekt, kan sannsynlighetene erstattes med fordelingen av de ukjente sannsynlighetene $\mu_1, \mu_2 .. \mu_6$. Igjen er hovedpoenget at innenfor hver krysningstype vil de forskjellige triadetypene være like vanlige.

Det er viktig å merke seg at triadetypen 111 har en frekvens som multipliseres med 2 i forhold til typene 112 og 110. Dette skyldes at typen egentlig består av to like hyppige undertyper: den typen som har *A*-allelet fra mor og den typen som har *A*-allelet fra far. Ved vanlig genotyping vil man ikke være i stand til å skille disse. Dersom det teknisk er mulig å si om et heterozygot barn har fått *A*-allelet fra mor eller fra far, vil det kunne skilles mellom 16 forskjellige triadetyper og dette kan inkorporeres i metodene som faktisk da blir enklere. Spesielt blir det da lettere å studere genomisk imprinting (at allelet *A* har en annen effekt når det kommer fra far enn når det kommer fra mor).

EN LOG-LINEÆR MODELL FOR ANALYSE AV EFFEKT AV BARNETS (PASIENTENS) ALLELER

Dersom det er sammenheng mellom forekomst av allelet *A* hos barnet og risiko for sykdom, vil fordelingen av triadetyper blant pasient-triadene være endret i forhold til fordelingen i befolkningen (tabell 4). I tabell 5 er forventet fordeling av pasient-triadene vist for en sykdom hvor barn som er heterozygote (har kun et *A*-allel) har en R_1 ganger økt risiko for å bli syke, mens barn som har to *A*-alleler har en R_2 ganger økt risiko for å bli syke. Ved hjelp av Bayes formel kan det vises at forekomsten av triadetyper blant pasient-triadene blir forskjøvet i forhold til fordelingen i befolkningen som vist i tabell 5 (3). Parametrene R_1 og R_2 kan derfor tolkes som relativ risiko, og kan estimeres ved hjelp av Poisson regresjon. Denne metoden finnes i de fleste vanlige statistikkpakker. Forutsetningen er at programmet har innbygd mulighet for å definere en såkalt offset-variabel. De fleste programmer har denne muligheten.

Tabell 5. Fordeling av pasient-triader når det er mulig effekt av at allelet *A* finnes hos barnet.

M F B[1]	KT[2]	Sannsynlighet (Ikke H-W[3])	Antall pasient-triader, $TGF\alpha$
2 2 2	1	$R_2\mu_1$	0
2 1 2	2	$R_2\mu_2$	0
2 1 1	2	$R_1\mu_2$	0
1 2 2	2	$R_2\mu_2$	1
1 2 1	2	$R_1\mu_2$	1
2 0 1	3	$R_1\mu_3$	3
0 2 1	3	$R_1\mu_3$	0
1 1 2	4	$R_2\mu_4$	1
1 1 1	4	$2R_1\mu_4$	5
1 1 0	4	μ_4	2
1 0 1	5	$R_1\mu_5$	9
1 0 0	5	μ_5	9
0 1 1	5	$R_1\mu_5$	16
0 1 0	5	μ_5	7
0 0 0	6	μ_6	67

[1] Kolonnen angir antall *A*-alleler hos henholdsvis mor, far og barn
[2] Krysningstyper av foreldre
[3] Ikke Hardy-Weinberg likevekt

Ved å ta logaritmen til frekvensen av hver triadetype, for eksempel typen 111, får man

$$log(\lambda_{111}) = log(2R_1\mu_4) = log(\mu_4) + log(2) + log(R_1)$$

Det er da relativt lett å se at frekvensene kan beskrives med en log-lineær modell med en indikatorvariabel for om barnet har ett *A*-allel og en for om barnet har to:

$$log(\lambda_{MFB}) = \gamma_i + log(2)I_{MFB=111} + \beta_1 I_{B=1} + \beta_2 I_{B=2}$$

Et lite triks gjør at man får inn tilleggskonstanten log(2) for triadetypen 111. Ved å definere en variabel som har verdien log(2) kun for typen 111 og har verdien 0 for alle andre triadetyper, og definere denne som *offset*-variabel i analysen, vil de fleste programmer for log-lineær Poisson-regresjon gi denne variabelen en konstant koeffisient med verdien 1. I stedet for å ha en indikatorvariabel for typen 111 og gi den en koeffisient på log(2), skjer altså det motsatte, men med samme resultat for modellen. Parametrene β_1 og β_2 er altså logaritmene til de relative risiki knyttet til henholdsvis ett og to *A*-allel sammenlignet med ingen *A*-allel. Logaritmene til $\mu_1, \mu_2,\ldots, \mu_6$ ($\gamma_1, \gamma_2,\ldots, \gamma_6$) inngår som konstantledd ved at modellen stratifiserer på krysningstypene. I praksis vil log(μ_1) estimeres som konstantleddet i modellen, og det som trengs for å tilpasse de andre, log(μ_2)…log(μ_6), vil være indikatorvariable for krysningstypene 2 til 6.

MODELL MED ANTAGELSE OM HARDY-WEINBERG LIKEVEKT

Analyser med modellen ovenfor er som nevnt stratifisert over de 6 krysningstypene. Dersom man er villig til å anta Hardy-Weinberg likevekt for de aktuelle allelene kan man spare noen parametre i modellen, og antagelig oppnå noe større statistisk styrke (4,17). Hardy-Weinberg frekvensene fra tabell 4 kan omskrives:

$$P(HW) = p^{M+F}(1-p)^{4-(M+F)} = [p/(1-p)]^{M+F}(1-p)^4$$

Dersom man tar logaritmene til disse sannsynlighetene, forenkles det hele til en lineær funksjon av summen (M+F) av antall *A*-allel hos mor og far:

$$log[P(HW)] = log[p/(1-p)](M+F) + 4log(1-p) = K_1(M+F)+K_2$$

Dette betyr at dersom stratifiseringen etter krysningstype erstattes av en kontinuerlig variabel som inneholder (M+F), kan den log-lineære modellen tilpasses med 4 parametre mindre:

$$log(\lambda_{MFB})=\alpha+K_1(M+F)+log(2)I_{MFB=111}+\beta_1 I_{B=1}+\beta_2 I_{B=2}$$

Koeffisienten K_1 for variabelen (M+F) vil være logit til allelfrekvensen p i befolkningen for allelet *A*. Det er litt overraskende at man faktisk kan estimere p ved å beregne anti-logit til koeffisienten K_1 i en modell for pasient-triader ($p = \exp(K_1)/(1+\exp(K_1))$. K_2 fra uttrykket over vil inngå i konstantleddet i modellen. Ved å sammenligne goodness-of-fit (–2log-likelihood) for denne modellen med modellen som stratifiserte på krysningstypene kan man faktisk også teste om allelene er i Hardy-Weinberg likevekt i befolkningen. Dette blir en χ^2-test med 4 frihetsgrader, litt avhengig av hvor mange konstantledd som estimeres. Testen er vist å ha begrenset statistisk styrke, og i og med at den kun baserer seg på om Hardy-Weinberg likevekt passer godt til dataene avhenger den også selvfølgelig av at modellen ellers er riktig spesifisert (4). Det er derfor naturlig å sjekke at fordelingen av genotyper blant foreldrene også ser ut til å være i Hardy-Weinberg likevekt før man baserer analysene på denne antagelsen.

ANALYSE AV EFFEKTER AV MORS ALLELER

Da den log-lineære modellen ble utviklet var hovedpoenget å finne en metode som gjorde det mulig å studere effekten av alleler hos mor. Mors inntak av enkelte vitaminer og medikamenter i begynnelsen av svangerskapet har betydning for risikoen for at barnet blir født med misdannelser. Det kan derfor godt tenkes at gener som regulerer mors metabolisme av vitaminer eller medikamenter påvirker risikoen for barnet. Et eksempel på et slikt gen et det såkalte MTHFR-genet som har sammenheng med metabolisme av vitaminet folat (18).

Den store fordelen ved å betrakte alle triadetypene i tabell 4 og 5 er at effekter av alleler hos mor, men ikke nødvendigvis hos barnet studeres direkte i den log-lineære modellen. I tabell 6 fremstilles fordelingen av de 15 triadetypene når det tenkes effekt både av barns og mors alleler. S_1 og S_2 er her de relative risikoene knyttet til at mor har henholdsvis ett eller to *A*-alleler. Estimering av log av alle fire relativ risikoparametre kan nå gjøres med å utvide den log-lineære modellen:

$$log(\lambda_{MFB}) = \gamma_i + log(2)I_{MFB=111} + \beta_1 I_{B=1} + \beta_2 I_{B=2} + \beta_3 I_{M=1} + \beta_4 I_{M=2}$$

Denne modellen kan også tilpasses med færre parametre ved å anta Hardy-Weinberg likevekt på samme måte som ovenfor.

Tabell 6. Fordeling av pasient-triader når det er mulig effekt av at allelet *A* finnes hos barnet og mulig effekt av at allelet finnes hos mor.

M F B[1]	KT[2]	Sannsynlighet (Ikke H-W[3])	Antall pasient-triader, *TGFα*
2 2 2	1	$R_2S_2\mu_1$	0
2 1 2	2	$R_2S_2\mu_2$	0
2 1 1	2	$R_1S_2\mu_2$	0
1 2 2	2	$R_2S_1\mu_2$	1
1 2 1	2	$R_1S_1\mu_2$	1
2 0 1	3	$R_1S_2\mu_3$	3
0 2 1	3	$R_1\mu_3$	0
1 1 2	4	$R_2S_1\mu_4$	1
1 1 1	4	$2R_1S_1\mu_4$	5
1 1 0	4	$S_1\mu_4$	2
1 0 1	5	$R_1S_1\mu_5$	9
1 0 0	5	$S_1\mu_5$	9
0 1 1	5	$R_1\mu_5$	16
0 1 0	5	μ_5	7
0 0 0	6	μ_6	67

[1] Kolonnen angir antall *A*-alleler hos henholdsvis mor, far og barn
[2] Krysningstyper av foreldre
[3] Ikke Hardy-Weinberg likevekt

Som epidemiolog skulle man tro at effektene av mors og barns alleler var korrelerte på en slik måte at de gav gjensidig effektforveksling ("confounding"). Underlig nok skjer ikke det i disse analysene. Når det stratifiseres på krysningstype blir effektene uavhengige av hverandre, og effekten av barnets alleler endrer seg ikke avhengig av om man "justerer" for effekten av mors alleler.

Denne modellen tilpasser effektene av ett og to alleler separat. Den samlede signifikansen av en effekt av barnets alleler finnes ved å tilpasse modellen over uten de to leddene for effekt av barnets alleler. Forskjellen i –2*log-likelihood mellom de to modellene gir en χ^2 med to frihetsgrader for samlet test av effekt av de to parameterne. Tilsvarende testes effekt av de to parameterne for effekt av mors alleler.

Alternativt til å tilpasse modellen med uavhengig estimering av effekten av et og to alleler, kan man tilpasse modellen med en parameter for enten å estimere en recessiv effekt, en gen-dose effekt eller en dominant effekt. En recessiv effekt fremkommer ved å sløyfe leddet med effekt av et allel. En gen-dose effekt fremkommer ved å bruke en variabel som inneholder antall alleler hos mor eller barn som en kontinuerlig variabel. For å konstruere en dominant effekt må man bruke en indikatorvariabel for at mor eller barn har minst et allel.

Testene for allel-effekter basert på denne modellen har overraskende god statistisk styrke. Med en blandet populasjon med frekvenser for *A*-allelet varierende mellom 10% og 30%, hadde metoden 70-80% styrke for å oppdage alleleffekter av størrelse RR = 2.5 kun basert på 100 pasient-triader (3).

PRAKTISK REGNEEKSEMPEL, ESTIMERING AV EFFEKTER AV MORS OG BARNS ALLELER

I tabell 7 er dataene våre for TGFα-genet tilrettelagt i en STATA-fil for analyse av effekter av mors og barns alleler. Variabelen *antall* angir antall triader av en bestemt triadetype. Når det ikke finnes triader av en bestemt type (type 1 her), bør antallet settes til "missing". Ikke alle programmer takler en null på samme måte. I GLIM for eksempel ville det gått greit å sette den til null. Programmet ville da selv oppfatte at gruppen måtte utelates. Triadetypen defineres av kombinasjonen av variablene *mor*, *far* og *barn*, som igjen angir antall *A*-alleler hos hver av de tre personene i en triade.

Når man så skal kjøre en Poisson-regresjon på disse dataene må *antall* angis som avhengig variabel. I STATA blir kommandolinjen for en samlet analyse av effektene både av barnets og morens alleler slik:

```
poisson antall t2-t6 m1 m2 b1 b2 , offset(set)
```

Indikatorvariablene t2 til t6 vil tilpasse frekvensene av krysningstypene med en serie konstantledd. Koeffisientene til variablene b1 og b2 vil være $\log(R_1)$ og $\log(R_2)$, og koeffisientene til variablene m1 og m2 vil være $\log(S_1)$ og $\log(S_2)$. Utskriften av kjøringen kan sees i tabell 8.

For å kjøre en tilsvarende analyse som forutsetter Hardy-Weinberg likevekt i befolkningen benyttes variabelen MF i stedet for variablene t2 til t6 i STATA-kommandoen. Kommandoen og uskriften fra STATA ved kjøring av kommandoen nedenfor kan sees i tabell 9.

```
poisson antall MF m1 m2 b1 b2 , offset(set)
```

Det første interessante fra disse to kjøringene er å se på om antagelsen om Hardy-Weinberg likevekt virker rimelig. Differansen i –2*log-likelihood for de to modellene er 2,19. Dette kan oppfattes som en χ^2 observator med 3 frihetsgrader i dette tilfellet. Det er klart at dette ikke er signifikant, og at modellen som antar Hardy-Weinberg passer nesten like godt til dataene som den uten denne antagelsen. Det er allikevel et spørsmål om en ikke-signifikant χ^2 alltid vil være god nok begrunnelse til å kunne bruke en analyse som antar Hardy-Weinberg likevekt når antagelsen egentlig

Tabell 7. STATA-datafil med variabler for analyse av *TGFα*.

mor	far	barn	type	MF	log(2)	antall	t2	t3	t4	t5	t6	m1	m2	b1	b2
2	2	2	1	4	0	.	0	0	0	0	0	0	1	0	1
2	1	2	2	3	0	0	1	0	0	0	0	0	1	0	1
2	1	1	2	3	0	0	1	0	0	0	0	0	1	1	0
1	2	2	2	3	0	1	1	0	0	0	0	1	0	0	1
1	2	1	2	3	0	1	1	0	0	0	0	1	0	1	0
2	0	1	3	2	0	3	0	1	0	0	0	0	1	1	0
0	2	1	3	2	0	0	0	1	0	0	0	0	0	1	0
1	1	2	4	2	0	1	0	0	1	0	0	1	0	0	1
1	1	1	4	2	0,693	5	0	0	1	0	0	1	0	1	0
1	1	0	4	2	0	2	0	0	1	0	0	1	0	0	0
1	0	1	5	1	0	9	0	0	0	1	0	1	0	1	0
1	0	0	5	1	0	9	0	0	0	1	0	1	0	0	0
0	1	1	5	1	0	16	0	0	0	1	0	0	0	1	0
0	1	0	5	1	0	7	0	0	0	1	0	0	0	0	0
0	0	0	6	0	0	67	0	0	0	0	1	0	0	0	0

ikke er nødvendig men ofte vil senke p-verdiene. Uansett hvordan man utfører analysene er i dette tilfellet ikke noen av parameterne for alleleffekter signifikante. Det er heller ikke noen tendens i parameterestimatene som kunne tyde på at det var snakk om dominante, recessive eller gen-dose effekter av allelene hos mor og hos barn.

TEST FOR IMPRINTING

Selv om man ikke finner hovedeffekter av alleler i analysene ovenfor kan det være relevant å teste om det er tegn til genomisk imprinting (1). Dette ville innebære at *A*-allelet hos barnet ville ha en annen effekt når det kom fra far enn når det kom fra mor. Det finnes opplagt informasjon i fordelingen av triadetyper om imprinting. For eksempel kan man sammenligne frekvensen av triadetypen 101 (tabell 4), hvor barnet får allel *A* fra mor, med triadetypen 011, hvor *A* kommer fra far.

En robust og enkel test for imprinting er utviklet av Weinberg (6). Den store fordelen med testen er at også den kan utføres ved hjelp av vanlig statistisk programvare, i dette tilfellet et program for logistisk regresjon. I tillegg estimeres effekten av alleler hos mor. Dersom det er snakk om imprinting må disse effektene justeres gjensidig. Effekt av mors alleler kan forveksles med imprinting og omvendt.

Tabell 8. Utskrift av STATA-kjøring av log-lineær Poisson regresjonsanalyse av dataene i tabell 7 for å estimere alleleffektene av *TGFa* uten å anta Hardy-Weinberg likevekt.

```
Poisson regression                      Number of obs = 14
                                        LR chi2(8)    = 262.62
                                        Prob > chi2   = 0.0000
Log likelihood = 23.418635              Pseudo R2     = 0.8486
```

Antall	Coef.	Std. Err.	z	P>z	[95% Conf.	Interval]
t2	-5.182211	1.01299	-5.12	0.000	-7.167634	-3.196788
t3	-4.405169	.8555896	-5.15	0.000	-6.082094	-2.728245
t4	-3.572193	.5523294	-6.47	0.000	-4.654739	-2.489647
t5	-2.044587	.3047852	-6.71	0.000	-2.641954	-1.447219
m1	-.1235253	.3086472	-0.40	0.689	-.7284627	.4814122
m2	.3564279	.921327	0.39	0.699	-1.44934	2.162196
b1	.4119311	.3001951	1.37	0.170	-.1764405	1.000303
b2	-.2364413	.8407562	-0.28	0.779	-1.884293	1.411411
cons	4.204693	.1221694	34.42	0.000	3.965245	4.44414
set	(offset)					

Tabell 9. Utskrift av STATA-kjøring av log-lineær Poisson regresjonsanalyse av dataene i tabell 7 for å estimere alleleffektene av *TGFa*, antagelse om Hardy-Weinberg likevekt.

```
Poisson regression                      Number of obs = 14
                                        LR chi2(5)    = 260.51
                                        Prob > chi2   = 0.0000
Log likelihood = -24.470779             Pseudo R2     = 0.8418
```

Antall	Coef.	Std. Err.	z	P>z	[95% Conf.	Interval]
MF	-1.901866	.2316401	-8.21	0.000	-2.355873	-1.44786
m1	-.0867237	.2796593	-0.31	0.756	-.6348459	.4613985
m2	.1756417	.6841549	0.26	0.797	-1.165277	1.516561
b1	.3110739	.2827483	1.10	0.271	-.2431026	.8652503
b2	.1132124	.8136223	0.14	0.889	-1.481458	1.707883
cons	4.175212	.1197443	34.87	0.000	3.940518	4.409907
set	(offset)					

Tabell 10. Fordeling av pasient-triader når det er mulig effekt både av at allelet *A* finnes hos barnet (R_p og R_2) og av at allelet finnes hos mor (S_1 og S_2). I tillegg har et enkelt *A*-allel en forskjellig effekt hos barnet når det kommer fra mor ($J_M R_p$) i forhold til om det kommer fra far (R_p).

M F B[1]	KT[2]	Sannsynlighet (Ikke H-W[3])	Antall pasient-triader,TGFα
2 2 2	1	$S_2R_2\mu_1$	0
2 1 2	2	$S_2R_2\mu_2$	0
2 1 1	2	$J_MS_2R_p\mu_2$	0
1 2 2	2	$S_1R_2\mu_2$	1
1 2 1	2	$S_1R_p\mu_2$	1
2 0 1	3	$J_MS_2R_p\mu_3$	3
0 2 1	3	$R_p\mu_3$	0
1 1 2	4	$S_1R_2\mu_4$	1
1 1 1	4	$S_1(1+J_M)R_p\mu_4$	5
1 1 0	4	$S_1\mu_4$	2
1 0 1	5	$J_M S_1R_p\mu_5$	9
1 0 0	5	$S_1\mu_5$	9
0 1 1	5	$R_p\mu_5$	16
0 1 0	5	μ_5	7
0 0 0	6	μ_6	67

[1] Kolonnen angir antall *A*-alleler hos henholdsvis mor, far og barn
[2] Krysningstyper av foreldre
[3] Ikke Hardy-Weinberg likevekt

I tabell 10 er de hypotetiske frekvensene for triadetypene vist i en situasjon hvor det er tenkt imprinting. R_2 er relativ risiko for barn som har to *A*-alleler, R_p er relativ risiko for barn som kun har et *A*-allel fra far, og J_MR_p er relativ risiko for barn som kun har et *A*-allel fra mor. J_M er altså overrisikoen for et allel som kommer fra mor. J_M vil være mindre enn 1 dersom allelet fra far har større effekt. På grunn av triadegruppen 111 som er en blanding av barn som har et A-allel fra mor og fra far, kan ikke J_M estimeres ved en enkel utvidelse av Poisson-regresjonsmodellen. I stedet kan man se på forholdet mellom triadetyper direkte på en måte som kan skille ut J_M slik at den kan estimeres med bidrag fra flest mulig triadetyper. Dette er utgangspunktet for Weinbergs metode (6, table 3).

Innenfor krysningstype 2 kan man estimere J_M som en odds ratio. Oddsen for type 212 i forhold til 122 er S_2/S_1. Videre er oddsen for type 211 i forhold til 121 J_MS_2/S_1. Altså er J_M odds ratio mellom disse fire gruppene. Tilsvarende fremkommer J_M som en odds ratio innenfor krysningstype 5. For krysningstype 3 er oddsen mellom gruppen 201 og 021 J_MS_2. Krysningstypene 1, 4 og 6 benyttes ikke. Det må defineres en avhengig variabel (dep) som angir teller-gruppen for de 6 odds forholdene og en variabel som angir om barnet har et *A*-allel. For samtidig å estimere S_2 og S_1 trengs i tillegg en variabel som angir om mor og far til sammen har mer enn et *A*-allel (krysningstypene 2 og 3) samt en variabel som angir differansen mellom en indikatorvariabel for at mor og far til sammen har et *A*-allel (krysningstype 5) og en indikatorvariabel for at mor og far til sammen har mer enn 2 *A*-alleler (krysningstype 2). Disse variablene inngår så i en logistisk regresjon uten konstantledd:

$$log[P(\text{"A fra mor"})/P(\text{"A ikke fra mor"})] = \beta_1 I_{B=1} + \beta_2 I_{M+F>1} + \beta_3 (I_{M+F=1} - I_{M+F>2})$$

Hele modellen er betinget eller stratifisert på krysningstype og på antall A-alleler barnet har. Dette fjerner parametrene R_p og R_2 fra modellen. Tolkningen av parametrene er at $\exp(\beta_1)$ er J_M, $\exp(\beta_2)$ er S_2 og $\exp(\beta_3)$ er S_1. Dersom det virkelig er tendens til imprinting for et aktuelt gen, vil estimeringen av effekten av mors alleler i den tidligere beskrevne Poisson-modellen uten imprinting bli feilestimert. Effektene må derfor estimeres i en analyse som tillater imprinting. Tilsvarende må også en analyse av imprinting justeres for mulig effekt av mors alleler. Denne feilkilden er det så vidt vi vet bare denne metoden som justerer for (6).

I tabell 11 er det verifisert at den foreslåtte modellen virkelig gir estimering av parameterne J_M, S_2 og S_1. Variablene som brukes for å "trekke ut" parameterne har egentlig ikke noen spesiell tolkning.

EKSEMPEL PÅ TEST FOR IMPRINTING

I tabell 12 er dataene våre tilrettelagt for analyse av imprinting i en STATA-fil. De fire hjelpevariablene dep (avhengig variabel), Ib (imprinting-variabelen), m2 (triadetyper med M+F>1) og m1 ($I_{M+F=1} - I_{M+F>2}$). Variablene har manglende verdier for de triadetypene som ikke inngår i analysen.

I analysen må *antall* angies som en variabel som angir vekt til gruppen, eller hvor mange enheter av den aktuelle triadetypen som finnes i dataene. De fleste program for logistisk regresjon tillater at man kan fjerne konstantleddet. I STATA gjøres det ved at man legger til opsjonen *nocon* i kommandolinjen:

```
logit dep Ib m2 m1 [fweight = antall] , nocon
```

STATA-utskriften av imprintinganalysen kan sees i tabell 13. Det er klart at det her er ingen signifikant tendens verken til imprinting eller til effekt av mors alleler. Estimatet av imprinting-effekten er J_M = exp(–0,8361) = 0,43 (95% KI: 0,12–1,51), S_2 = exp(1,2339) = 3,43 (95% KI: 0,36–32,54) og S_1 = exp(0,3806) = 1,46 (95% KI: 0,55–3,87). Tendensen er altså at et *A*-allel har dobbelt så stor effekt når det kommer fra far, og ikke fra mor (RR = 0,43). At mor selv har to *A*-alleler øker risikoen 3,43 ganger, mens risikoen øker 1,46 ganger dersom mor er heterozygot. Ingen av disse effektene er altså signifikante. Konfidensintervall for disse relative risikoene fremkommer ved å eksponensialtransformere intervallene for koeffisientene nedenfor. Igjen er det viktig å huske at det er relative risikoer som estimeres, ikke odds ratioer slik vi er vant til ved logistisk regresjon. Dette sees best ved å gå tilbake til tabell 10 og se hvordan størrelsene J_M , S_1 og S_2 er definert.

INTERAKSJON MELLOM GEN OG MILJØ

Til nå har utledningene og beregningene kun dreiet seg om å estimere effekter av alleler. For mange epidemiologer er det overraskende at pasient-triader kan inneholde så mye informasjon om disse effektene. Det er ikke mindre overraskende at pasient-triadene faktisk gir mulighet til å estimere interaksjon mellom et gen og miljøfaktorer. Prinsippet er imidlertid nokså enkelt når man først har innsett at effekter av alleler kan estimeres i form av relativ risiko fra triadedata (5). Dersom man har en dikotom eksponeringsvariabel som skiller mellom eksponerte og ikke-eksponerte, kan man også enkelt skille eksponerte triader fra ueksponerte. Deretter kan man estimere de relative risikoer knyttet til allelene for hver gruppe separat. Ratioene av alleleffektene mellom eksponert og ueksponert gruppe vil direkte måle grad av interaksjon. Et eksempel på interaksjon kan være at allelet ikke har effekt blant de ueksponerte, mens det har en viss effekt blant de eksponerte.

For at ikke Poisson-modellen skal bli for komplisert vil vi anta at *A*-allelet har en dominant effekt. Dette reduserer antall effekt-parametre til det halve. Dersom I_E er en indikatorvariabel for eksponering, kan modellen for gen-miljø interaksjon skrives som følger:

$$log(\lambda_{MFB}) = \gamma_i + log(2) I_{MFB=111} + \beta_1 I_{B\geq 1} + \beta_2 I_{M\geq 1} + \gamma_i' I_E + \beta_3 I_{B\geq 1} I_E + \beta_4 I_{M\geq 1} I_E$$

Tabell 11. Verifisering av at parametrene J_M, S_2 og S_1 fra imprintingmodellen estimeres med varlablene Ib, m2 og m1. Logaritmen til odds-forholdet mellom de angitte triadetypene svarer til log(odds) fra regresjonsmodellen.

Triade-type	Mor	Far	Barn	Frekvens	Odds-forhold	*dep*	*Ib*	*m2*	*m1*	log(odds) fra regresjonsmodellen: $log(J_M)Ib + log(S_2)m2 + log(S_1)m1$		
2	2	1	2	$S_2R_2\mu_2$	S_2/S_1	1	0	1	-1		$log(S_2)$	$-log(S_1)$
4	1	2	2	$S_1R_2\mu_2$		0	0	1	-1			
3	2	1	1	$J_MS_2R_p\mu_2$	J_MS_2/S_1	1	1	1	-1	$log(J_M)$	$+ log(S_2)$	$- log(S_1)$
5	1	2	1	$S_1R_p\mu_2$		0	1	1	-1			
6	2	0	1	$J_MS_2R_p\mu_3$	J_MS_2	1	1	1	0	$log(J_M)$	$+ log(S_2)$	
7	0	2	1	$R_p\mu_3$		0	1	1	0			
11	1	0	1	$J_MS_1R_p\mu_5$	J_MS_1	1	1	0	1	$log(J_M)$		$+ log(S_1)$
13	0	1	1	$R_p\mu_5$		0	1	0	1			
12	1	0	0	$S_1\mu_5$	S_1	1	0	0	1			$+ log(S_1)$
14	0	1	0	μ_5		0	0	0	1			

Parametrene γ'_i vil i dette tilfellet tilpasse Poisson-modellen til fordelingen av krysningstyper i den eksponerte gruppen. Effekten av barnets og morens *A*-allel estimeres av β_1 og β_2 for de ueksponerte, mens tilleggseffektene blant de eksponerte estimeres av β_3 og β_4 . Dersom man heller vil estimere effekten av allelet separat for de eksponerte og de ueksponerte vil modellen se slik ut:

$$log(\lambda_{MFB}) = \gamma_i + log(2)I_{MFB=111} + \beta_1 I_{B\geq 1}(1 - I_E) + \beta_2 I_{M\geq 1}(1 - I_E) + \gamma_i' I_E + \beta_3 I_{B\geq 1} I_E + \beta_4 I_{M\geq 1} I_E$$

Tabell 12. STATA-datafil med variabler for imprintinganalyse av *TGFα*.

mor	far	barn	type	antall	dep	lb	m2	m1
2	2	2	1	.	.	.	.	.
2	1	2	2	0	1	0	1	-1
2	1	1	2	0	1	1	1	-1
1	2	2	2	1	0	0	1	-1
1	2	1	2	1	0	1	1	-1
2	0	1	3	3	1	1	1	0
0	2	1	3	0	0	1	1	0
1	1	2	4	1	.	.	.	.
1	1	1	4	5	.	.	.	.
1	1	0	4	2	.	.	.	.
1	0	1	5	9	1	1	0	1
1	0	0	5	9	1	0	0	1
0	1	1	5	16	0	1	0	1
0	1	0	5	7	0	0	0	1
0	0	0	6	67	.	.	.	.

Med litt kyndighet i bruk av regresjonsmodeller kan man tilpasse modellen til mange forskjellige former for allel-effekter og samspill. Det er selvfølgelig en grunnleggende begrensning for disse modeller basert på pasient-triader at man ikke kan estimere hoved-effekter av eksponering. For å kunne gjøre det må man antagelig ha en kontrollgruppe.

EKSEMPEL PÅ GEN-MILJØ INTERAKSJON

Det finnes i litteraturen en del studier av samspill mellom TGFα og mors røking (19,20). Disse studiene er basert på pasient-kontrolldata og ikke triader. I våre data for leppe- og ganespalte finnes opplysninger om røking. Vi har derfor muligheten til å skille triadene som er blitt analysert ovenfor i to grupper, de hvor mor har røkt, og de hvor mor ikke har røkt. I tabell 14 er disse dataene tilrettelagt som en STATA-fil.

Tabell 13. Utskrift av STATA-kjøring av logistisk regresjonsanalyse av dataene i tabell 11 for å estimene effekten av mors alleler sammen med imprintingeffekt for *TGFa*.

```
Logit estimates                   Number of obs = 46
                                  LR chi2(3)    =  .
Log likelihood = -30.826671       Prob > chi2   =  .

Dep      Coef.    Std. Err.    z    P>z    [95% Conf. Interval]

Ic     -.8361171  .6377422  -1.31  0.190   -2.086069   .4138347
Ip1    1.233884   1.147303   1.08  0.282   -1.014789   3.482558
Idiff  .3806017   .4965792   0.77  0.443   -.5926758   1.353879
```

Tabell 14. STATA-datafil med variabler for analyse av *TGFa* og mors røking.

mor	far	barn	type	set	antall	t2	t3	t4	t5	t6	t2e	t3e	t4e	t5e	T6e	md	bd	mdie	bdie	mdn	bdn	sm
2	2	2	1	0		0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1	0
2	1	2	2	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1	0
2	1	1	2	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1	0
1	2	2	2	0	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1	0
1	2	1	2	0	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1	0
2	0	1	3	0	2	0	1	0	0	0	0	0	0	0	0	1	1	0	0	1	1	0
0	2	1	3	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0
1	1	2	4	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0	1	1	0
1	1	1	4	0,693	2	0	0	1	0	0	0	0	0	0	0	1	1	0	0	1	1	0
1	1	0	4	0	1	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0
1	0	1	5	0	5	0	0	0	1	0	0	0	0	0	0	1	1	0	0	1	1	0
1	0	0	5	0	7	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	0
0	1	1	5	0	8	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0
0	1	0	5	0	6	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	6	0	42	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
2	2	2	1	0		0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	1
2	1	2	2	0	0	1	0	0	0	0	1	0	0	0	0	1	1	1	1	0	0	1
2	1	1	2	0	0	1	0	0	0	0	1	0	0	0	0	1	1	1	1	0	0	1
1	2	2	2	0	0	1	0	0	0	0	1	0	0	0	0	1	1	1	1	0	0	1
1	2	1	2	0	0	1	0	0	0	0	1	0	0	0	0	1	1	1	1	0	0	1
2	0	1	3	0	1	0	1	0	0	0	0	1	0	0	0	1	1	1	1	0	0	1
0	2	1	3	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	1	0	0	1
1	1	2	4	0	1	0	0	1	0	0	0	0	1	0	0	1	1	1	1	0	0	1
1	1	1	4	0,693	3	0	0	1	0	0	0	0	1	0	0	1	1	1	1	0	0	1
1	1	0	4	0	1	0	0	1	0	0	0	0	1	0	0	1	0	1	0	0	0	1
1	0	1	5	0	4	0	0	0	1	0	0	0	0	1	0	1	1	1	1	0	0	1
1	0	0	5	0	2	0	0	0	1	0	0	0	0	1	0	1	0	1	0	0	0	1
0	1	1	5	0	8	0	0	0	1	0	0	0	0	1	0	0	1	0	1	0	0	1
0	1	0	5	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1
0	0	0	6	0	25	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1

Det finnes mange måter å kjøre interaksjonsanalysen på. Nedenfor er gitt to eksempler på STATA-kommandolinjer som utfører Poisson-regresjonen på litt forskjellige måter. I den første vil koeffisientene til variablene *md* og *bd* estimere baseline-effektene av A-allelet hos mor og barn. Koeffisientene til variablene *mdie* og *bdie* vil estimere tilleggseffekten av allelet for de røke-eksponerte. I den neste kommandoen estimeres effektene av allelet separat for de eksponerte og de ueksponerte. Dette svarer til de to formlene ovenfor.

```
poisson antall t2-t6 t2e-t6e md mdie bd bdie , offset(set)

poisson antall t2-t6 t2e-t6e mdn mdie bdn bdie , offset(set)
```

I tabell 15 sees utskriften fra STATA for de to analysene. Legg merke til at den første analysen viser at det ikke er noen signifikant interaksjon. Det nærmeste er en p-verdi på 0,074 for en interaksjon mellom barnets allel og mors røking.

I neste uskrift i tabell 15 fremkommer det imidlertid at det er en signifikant effekt av *A*-allelet i røkegruppen (RR = 3,24, 95% KI: 1,07–9,83, p = 0,038), men ikke blant ikke-røkere. Her er analysene gjort kun som en illustrasjon og med en blandet gruppe av både leppe- og ganespaltepasienter. De kan derfor ikke tillegges mye vekt. Men dette er selvfølgelig en interaksjon som SAM-prosjektet vil undersøke i et større materiale og med mer veldefinerte pasientgrupper.

Som en sidekommentar bør det nevnes at denne interaksjonen egentlig er ganske interessant. Hypotesen om en interaksjon går ut på at når mor røker oppstår det en viss grad av hypoksi (surstoffmangel) hos fosteret. Når dette kombineres med at et foster har et mutert allel for et vekstfaktorgen (TGFα) som er aktivt i lukkingen av leppen, kan det tenkes at kombinasjonen gir en målbar øket risiko. Det skal også nevnes at røking i seg selv er en nokså sikker risikofaktor for leppe- og ganespalte med en relativ risiko på vel 1,5. Interaksjonen dukket først opp ved at Hwang et al. så den i et pasient-kontroll materiale fra Maryland (19). I 1999 publiserte imidlertid Christensen et al. en god pasient-kontroll studie som ikke fant denne interaksjonen i et materiale fra Danmark (20).

Triader kan også brukes til å se på samspill mellom alleler av forskjellige gen. Det forutsetter at triadene klassifiseres i triadetyper for begge genene simultant. I praksis vil det da være like enkelt å benytte en fil med en triade pr. linje som en aggregert fil. Den aggregerte filen vil iallfall ha 15^2 linjer.

FEILKILDER I ANALYSENE

Utgangspunktet for innføringen av triadedesignet var de mulige feilkildene knyttet til befolkningsstratifisering i en vanlig pasient-kontrollstudie. Det er imidlertid viktig å være klar over at pasient-triadedesignet også har en del mulige feilkilder. Den mest grunnleggende knytter seg til at allelene som studeres antas å bli overført ved vanlig Mendelsk arv i befolkningen, altså at begge alleler har like stor sannsynlighet for å bli overført fra foreldre til barn. Dersom et allel hos barnet for eksempel er knyttet til risiko for intrauterin død, vil allelet være underrepresentert blant alle nyfødte i forhold til forventet fordeling ut fra foreldrene. Dette kan dukke opp som en falsk effekt i pasientgruppen, eller maskere en reell effekt. Dersom man finner en effekt av et allel vil man ofte være interessert i å kunne sjekke at det ikke finnes tendens til en tilsvarende effekt blant kontroll-triader hvor barna ikke er pasienter. Alternativt kan man konstruere kontroll-triader av foreldrene til pasienten og eventuelle søsken av pasienten (4,21).

Ofte vil man ønske å benytte triadedesignet på medfødte tilstander som i seg selv godt kan tenkes å være forbundet med redusert intrauterin overlevelse. Dersom overlevelsen til et foster som har fått en misdannelse ikke er avhengig av allelene som studeres, kun av selve tilstanden (fenotypen), skal allikevel triadedesignet estimere allel-effektene korrekt (3).

Tabell 15. Utskrift av STATA-analyse av *TGFa* og røking.

```
Poisson regression                              Number of obs = 28
                                                LR chi2(13)   = 276.23
                                                Prob > chi2   = 0.0000
Log likelihood = -34.601071                     Pseudo R2     = 0.7997
```

antall	Coef.	Std. Err.	z	P>z	[95% Conf.	Interval]
t2	-.3964819	.9179554	-0.43	0.666	-2.195641	1.402678
t3	.2966447	.937212	0.32	0.752	-1.540257	2.133546
t5	2.150614	.6453686	3.33	0.001	.885715	3.415513
t6	3.998954	.7603558	5.26	0.000	2.508685	5.489224
t2e	-15.8696	1221.404	-0.01	0.990	-2409.778	2378.039
t3e	-1.787603	1.430329	-1.25	0.211	-4.590996	1.015789
t4e	-.2512466	1.12762	-0.22	0.824	-2.461341	1.958848
t5e	-1.202192	.6407805	-1.88	0.061	-2.458099	.0537142
t6e	-.5188124	.2526046	-2.05	0.040	-1.013908	-.0237165
md	-4.78e-17	.3779622	-0.00	1.000	-.7407923	.7407923
mdie	-.2513592	.6299354	-0.40	0.690	-1.48601	.9832916
bd	-.0353784	.3760766	-0.09	0.925	-.7724751	.7017182
bdie	1.212277	.6793648	1.78	0.074	-.1192535	2.543807
cons	-.2612663	.7445346	-0.35	0.726	-1.720527	1.197995
set	(offset)					

```
Poisson regression                              Number of obs = 28
                                                LR chi2(13)   = 276.23
                                                Prob > chi2   = 0.0000
Log likelihood = -34.601071                     Pseudo R2     = 0.7997
```

antall	Coef.	Std. Err.	z	P>z	[95% Conf.	Interval]
t2	-.6931266	1.017695	-0.68	0.496	-2.687773	1.301519
t4	-.2966447	.937212	-0.32	0.752	-2.133546	1.540257
t5	1.85397	.7583411	2.44	0.014	.3676482	3.340291
t6	3.70231	.8372315	4.42	0.000	2.061366	5.343253
t2e	-15.8696	1221.404	-0.01	0.990	-2409.778	2378.039
t3e	-1.787603	1.430329	-1.25	0.211	-4.590996	1.015789
t4e	-.2512466	1.12762	-0.22	0.824	-2.461341	1.958848
t5e	-1.202192	.6407805	-1.88	0.061	-2.458099	.0537142
t6e	-.5188124	.2526046	-2.05	0.040	-1.013908	-.0237165
mdn	-1.56e-17	.3779622	-0.00	1.000	-.7407923	.7407923
mdie	-.2513592	.5039476	-0.50	0.618	-1.239078	.73636
bdn	-.0353784	.3760766	-0.09	0.925	-.7724751	.7017182
bdie	1.176899	.5657763	2.08	0.038	.0679973	2.2858
cons	.0353784	.8228897	0.04	0.966	-1.577456	1.648213
set	(offset)					

Estimering av effektene av mors alleler er mer usikker enn tilsvarende estimering av effektene av barnets alleler. Her brukes egentlig fordelingen av alleler hos far som en slags kontroll. Dersom det er systematisk annerledes fordeling av alleler hos far enn hos mor, for eksempel i en befolkning hvor menn fra en etnisk gruppe systematisk velger partnere fra en annen etnisk gruppe men ikke omvendt, vil det kunne bli skjevheter. Derfor bør også effekter av mors alleler verifiseres ved at man viser at slike effekter ikke sees utenfor pasientgruppen. Weinberg og Umbach (22) har gitt en samlet gjennomgang av en rekke feilkilder ved forskjellige designtyper for studier av interaksjon mellom gen og miljø. Moralen er egentlig at det alltid er noe som kan gå galt.

Modellen som ligger til grunn for analysene her (tabell 10) spesifiserer at det til et bestemt allel er knyttet en bestemt overrisiko. Det innebærer egentlig at modellen oppfatter det aktuelle genet som et kandidatgen som virkelig har en etiologisk funksjon. Dersom man i stedet tenker på det aktuelle genet som en markør som kanskje er i koblingsulikevekt ("linkage disequilibrium") med et aktivt sykdomsgen i nærheten på samme kromosom blir modellen antagelig ikke helt riktig. Da vil det oppstå en grad av rekombinasjon mellom sykdomsgenet og markøren som modellen vår ikke tar hensyn til. Modellen bør allikevel kunne være egnet til å oppdage koblingsulikevekt, men de relative risikoestimatene bør da oppfattes litt løsere som et mål på grad av assosiasjon.

HJELP MED ARBEIDET

Forfatterne takker Lars-Christian Stene for nyttige kommentarer til manuskriptet.

REFERANSER

1. Barlow DP. Gametic imprinting in mammals. *Science* 1995; **270** (5242): 1610-3.
2. STATA Corporation. STATA User's Guide, Stata Press, Texas, USA, 2001.
3. Wilcox AJ, Weinberg CR, Lie RT. Distinguishing the effects of maternal and offspring genes through studies of case-parent triads. *Am J Epidemiol* 1998; **148**: 893-901.
4. Weinberg CR, Wilcox AJ, Lie RT. A log-linear approach to the case-parent-triad data: Assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting. *Am J Hum Genet* 1998; **62**: 969-78.
5. Umbach DM, Weinberg CR. The use of case-parent triads to study joint effects of genotype and exposure. *Am J Hum Genet* 2000; **66** (1): 251-61.
6. Weinberg CR. Methods for detection of parent-of-origin effects in genetic studies of case-parents triads. *Am J Hum Genet* 1999; **65** (1): 229-35.
7. Le Marchand L, Lum-Jones A, Saltzman B, Visaya V, Nomura AM, Kolonel LN. Feasibility of collecting buccal cell DNA by mail in a cohort study. *Cancer Epidemiol Biomarkers Prev* 2001; **10** (6): 701-3.
8. Garcia-Closas M, Egan KM, Abruzzo J, Newcomb PA, Titus-Ernstoff L, Franklin T, Bender PK, Beck JC, Le Marchand L, Lum A, Alavanja M, Hayes RB, Rutter J, Buetow K, Brinton LA, Rothman N. Collection of genomic DNA from adults in epidemiological studies by buccal cytobrush and mouthwash. *Cancer Epidemiol Biomarkers Prev* 2001; **10** (6): 687-96.
9. Feigelson HS, Rodriguez C, Robertson AS, Jacobs EJ, Calle EE, Reid YA, Thun MJ. Determinants of DNA yield and quality from buccal cell samples collected with mouthwash. *Cancer Epidemiol Biomarkers Prev* 2001; **10** (9): 1005-8.
10. Zheng S, Ma X, Buffler PA, Smith MT, Wiencke JK. Whole genome amplification increases the efficiency and validity of buccal cell genotyping in pediatric populations. *Cancer Epidemiol Biomarkers Prev* 2001; **10** (6): 697-700.
11. 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** (14): 1151-8.
12. Falk CT, Rubinstein P. Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. *Ann Hum Genet* 1987; **51**: 227-33.
13. Self SG, Longton G, Kopecky KJ, Liang KY. On estimating HLA/disease association with application to a study of aplastic anemia. *Biometrics* 1991; **47** (1): 53-61.
14. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* 1993; **52** (3): 506-16.
15. Cytel Sotfware Corporation. Statxact. Cambridge, USA, 1999.
16. Maestri NE, Beaty TH, Hetmanski J, Smith EA, McIntosh I, Wyszynski DF, Liang KY, Duffy DL, VanderKolk C. Application of transmission disequilibrium tests to nonsyndromic oral clefts: including candidate genes and environmental exposures in the models. *Am J Med Genet* 1997; **73** (3): 337-44.

17. Shields DC, Kirke PN, Mills JL, Ramsbottom D, Molloy AM, Burke H, Weir DG, Scott JM, Whitehead AS. The "thermolabile" variant of methylenetetrahydrofolate reductase and neural tube defects: An evaluation of genetic risk and the relative importance of the genotypes of the embryo and the mother. *Am J Hum Genet* 1999; **64** (4): 1045-55.
18. van der Put NM, Steegers-Theunissen RP, Frosst P, Trijbels FJ, Eskes TK, van den Heuvel LP, Mariman EC, den Heyer M, Rozen R, Blom HJ. Mutated methylenetetrahydrofolate reductase as a risk factor for spina bifida. *Lancet* 1995; **346** (8982): 1070-1.
19. Hwang SJ, Beaty TH, Panny SR, Street NA, Joseph JM, Gordon S, McIntosh I, Francomano CA. Association study of transforming growth factor alpha (TGF alpha) TaqI polymorphism and oral clefts: indication of gene-environment interaction in a population-based sample of infants with birth defects. *Am J Epidemiol* 1995; **141** (7): 629-36.
20. Christensen K, Olsen J, Norgaard-Pedersen B, Basso O, Stovring H, Milhollin-Johnson L, Murray JC. Oral clefts, transforming growth factor alpha gene variants, and maternal smoking: a population-based case-control study in Denmark, 1991-1994. *Am J Epidemiol* 1999; **149** (3): 248-55.
21. Spielman RS, Ewens WJ. The TDT and other family-based tests for linkage disequilibrium and association. *Am J Hum Genet* 1996; **59** (5): 983-9.
22. Weinberg CR, Umbach DM. Choosing a retrospective design to assess joint genetic and environmental contributions to risk. *Am J Epidemiol* 2000; **152** (3): 197-203.