

Cohort of Norway (CONOR):

Forskningspotensial, design og representativitet

Geir Aamodt ¹, Anne Johanne Sjøgaard ^{1,2}, Øyvind Næss ^{1,2}, Anne Cathrine Beckstrøm ¹, Sven Ove Samuelsen ^{1,3} og CONORs styringsgruppe

- 1) Divisjon for epidemiologi, Nasjonalt folkehelseinstitutt
- 2) Institutt for allmenn- og samfunnsmedisin, Universitetet i Oslo
- 3) Matematisk institutt, Universitetet i Oslo

Hovedpunkter

- CONOR er en sammenstilling av 10 helseundersøkelser med informasjon om helse, miljø og livsstil, enkelte selvrapporterte sykdommer, målinger av blodtrykk, og antropometriske mål, samt analyser av blodlipider
- Fullblod og DNA finnes lagret for deltakerne i CONOR Biobank
- CONOR er tilnærmet representative utvalg som i ulik grad avviker fra Norges befolkning
- CONOR passer til assosiasjonsstudier, men kan, ved å anvende standardisering, brukes til insidens- og prevalensstudier
- CONOR passer spesielt til årsaksforskning av sjeldne sykdommer med kompleks etiologi, slik som samspill mellom genetiske faktorer og miljø

Sammendrag

Sammen med Nasjonalt folkehelseinstitutt har universitetene i Tromsø, Trondheim, Bergen og Oslo investert i en sammenstilling av flere befolkningsbaserte helseundersøkelser til en stor database: Cohort of Norway (CONOR). Kohorten, som nå inkluderer om lag 180 000 personer med informasjon fra spørreskjemaer, målinger og blodprøver, er koblet til sykdomsregistre som Kreftregisteret og Dødsårsaksregisteret - og til sosioøkonomiske opplysninger fra NAV og SSB. Denne databasen er en verdifull kilde til epidemiologisk forskning. I helseundersøkelsene, som danner basis for CONOR, har man invitert utvalg med forskjellig sammensetning når det gjelder alder, kjønn og fødeland. CONOR kan derfor være beheftet med skjevheter. Spørsmålet er om dette er avgjørende for hvordan CONOR-data kan brukes?

I artikkelen diskuterer vi hvilke forskningsmuligheter CONOR-data gir. Først vil vi sammenlikne fordelingene av utvalgte demografiske variable, samt røyking og alkohol for hver av helseundersøkelsene, for CONOR samlet og for befolkningen i Norge. Deretter vil vi drøfte ulike epidemiologiske design som kan være aktuelle. Til slutt vil vi se på betingelser for å beregne antall kasus og kontroller som må inkluderes i studier av samspill mellom sykdom, gener og miljø. Slike studier gjør CONOR spesielt verdifull ettersom hel-genom assosiasjonsstudier har identifisert flere gener som er assosiert med sykdom. Nye undersøkelser av samspill mellom genetiske faktorer og miljø vil kreve store kohorter med informasjon om både genetikk og miljø.

Bakgrunn

CONOR er betegnelsen på en sammenstilling av helsedata og blodprøver, og er navnet på et nasjonalt samarbeid mellom Folkehelseinstituttet og universitetene i Oslo, Bergen, Trondheim og Tromsø. Totalt 10 helseundersøkelser, som ble gjennomført i tidsrommet 1994-2003, inngår i CONOR. Undersøkelsene er listet opp i tabell 1 med forkortelser og innsamlingsår. Data fra ytterligere tre undersøkelser vil bli lagt til når godkjenning fra Datatilsyn og REK foreligger. Dette gjelder Homocystin undersøkelsen i Hordaland (1992-93), den tredje helseundersøkelsen i Nord-Trøndelag (HUNT III 2006-08) og den sjette helseundersøkelsen i Tromsø, 2007-08). CONOR er koblet til Medisinsk fødselsregister, Dødsårsaksregisteret, Reseptregisteret, Kreftregisteret, NORHip (register over hoftebrudd i Norge 1994-2008), CVDNOR (register over hjerte- og karsykdom i Norge 1994-2009), Vannverksregisteret, benmassemålinger i NOREPOS (Norske epidemiologiske osteoporosestudier), en rekke tidligere helseundersøkelser, statistikk fra Arbeids- og velferdsetaten og sosio-demografiske registre i Statistisk sentralbyrå. For å få andre sykdomsendepunkter er koblinger til Norsk pasientregister aktuelle. CONORs historie og noe om materiale og metode, er publisert tidligere (1-4). Se også CONORs hjemmeside www.fhi.no/conor for ytterligere informasjon. CONOR er unik på grunn av størrelsen, og fordi man har EDTA-blod/DNA fra ca. 180 000 deltakere. Dette gir muligheter for studier av årsaker til en rekke sykdommer, der informasjon om miljøfaktorer og genetikk - og samspill mellom disse, kan inngå.

For problemstillinger der målet er å beskrive kildepopulasjonen skal et utvalg fra denne populasjonen ha de samme egenskaper som den populasjonen utvalget er hentet fra. For CONOR er utvalgsstrategien kompleks fordi man har invitert utvalg med forskjellig alder, kjønn og etnisk sammensetning til de ulike helseundersøkelsene. I tillegg er kun utvalgte fylker i Norge representert og fremmøteprosenten er forskjellig fra undersøkelse til

undersøkelse. Dette kan utgjøre enkelte begrensninger for hvordan en kohort som CONOR kan nyttes.

I denne artikkelen vil vi kort beskrive bakgrunnsdata fra de helseundersøkelsene som utgjør CONOR. Vi vil undersøke i hvilken grad helseundersøkelsene avviker fra befolkningen når det gjelder utvalgte bakgrunnsfaktorer, og vurdere om summen av helseundersøkelsene (CONOR) gir et riktig bilde av befolkningen eller ikke. Vi vil også diskutere hvilke studiedesign en slik sammensatt kohort kan brukes til.

Cohort of Norway

CONOR består av fem datakilder fra hver deltaker: Den *første* er vel 50 CONOR-spørsmål om livsstil (fysisk aktivitet, røyking, alkohol), skolegang og utdanning, bomiljø, familie og venner, arbeid og yrke, samt bruk av medisiner) hentet fra spørreskjemaer. I tillegg spørres det om helse og sykdom (for eksempel tidligere underarms- og lårhalsbrudd, symptomer og sykdommer innen hjerte/kar, hørsnue, allergi, muskel-skjelettsmerter og psykiske plager), samt sykdommer i familien. CONOR-spørsmålene er nærmere dokumentert på CONOR sin webside. Flere av spørsmålene er stilt litt forskjellig i de ulike helseundersøkelsene. De er imidlertid kodet om slik at de gir tilnærmet lik mening for alle helseundersøkelsene. Alle endringer, sammenslåinger og ulikheter er beskrevet i et eget dokument som følger utlevering av data fra databasen. Den *andre* kilden er målinger av blodtrykk, puls, høyde, vekt og livhoftemål, mens den *tredje kilden* er resultater fra analyser av glukose og lipider i serum tappet ved undersøkelsestidspunktet. Den *fjerde* kilden er EDTA-blod. Dette finnes nedfrosset i ”CONOR biobank”, lokalisert sammen med HUNT biobank i Levanger. DNA er en viktig *siste kilde* til data i CONOR, som er muliggjort takket være Norges forskningsråds satsing på funksjonell genomforskning gjennom forskningsprogrammet FUGE. Det er per oktober 2008

ekstrahert DNA fra ca 110 000 av full-blod prøvene, og flere vil bli ekstrahert på prosjektbasis. Dette materialet forvaltes og administreres av forskningssamarbeidet *Biohealth Norway* – som består av CONOR og Mor og barn- undersøkelsen. CONOR-data og biologisk materiale er tilgjengelig for forskere som oppfyller retningslinjer beskrevet på CONORs hjemmesider (www.fhi.no/conor).

Det er per november 2013 i alt 173 236 personer i CONOR. Videre er 7 309 personer fulgt opp i to undersøkelser, mens en person har deltatt ved tre undersøkelser. Etter at Homocystin undersøkelsen, HUNT III og Tromsø VI inkluderes i CONOR, vil antall personer være omlag 200 000. I forskningsprosjekter er søskengrupper, mor-barn par, far-barn par og mor-far-barn triader identifisert. Dette gir unike forskningsmuligheter for analyser av forholdet mellom arv og miljø, samt studier av forhold som strekker seg over flere generasjoner. Gjennom koblinger til Kreftregisteret (2006) og Dødsårsaksregisteret (2006) har vi funnet ut at det er 15 077 krefttilfeller i CONOR-materialet, og at 9 105 av deltakerne er døde (4).

CONOR og befolkningen i Norge

For å undersøke i hvilken grad CONOR-undersøkelsene samlet representerer befolkningen i Norge, har vi sammenliknet nøkkeltall for hver av helseundersøkelsene som er med i CONOR og CONOR samlet - med hele befolkningen. Vi har valgt variablene alder, kjønn, utdanning, fødeland, daglig røyking og bruk av alkohol. Verdiene for befolkningen er hentet fra Statistisk sentralbyrå for året 2000 (www.ssb.no) (alder, kjønn, røykevaner, utdanning og fødeland), mens informasjon om alkoholvaner er hentet fra spørreundersøkelser for året 2004 fra Statens institutt for rusmiddelforskning (www.sirus.no) (5).

Sosiodemografiske faktorer

I tabell 2 og 3 viser vi oppsummerende statistikk for de ulike helseundersøkelsene og for hele CONOR. I den siste kolonnen viser vi tilsvarende verdier for Norge. Vi har ekskludert personer under 20 år. Denne gruppen utgjør 0,2 % av CONOR deltakerne. Det er betydelige forskjeller i alderssammensetninger mellom undersøkelsene. For CONOR er det høyere andeler med 40-49, 60-69 og 70-79 åringer enn i befolkningen for øvrig, mens de yngste og eldste personene er underrepresentert (Tabell 2). I alle undersøkelser, med unntak av Innvandrere-HUBRO, har flere kvinner enn menn deltatt. I Oslo II var bare menn invitert. I CONOR er det totalt litt høyere andel kvinner enn i befolkningen som helhet (51,3 % versus 50,2 %).

Spørsmål om utdanning er stilt noe ulikt i de enkelte undersøkelsene, men er kodet om til en variabel som angir antall år med skole/utdanning totalt. For menn er gjennomsnittlig antall års utdanning for CONOR og for befolkningen henholdsvis 11,8 og 11,9 år (Tabell 2).

Tilsvarende verdier for kvinner er 11,6 år – både i CONOR og i befolkning. Antall års utdanning i befolkningen har vært relativt stabilt de siste ti årene.

Det er stor variasjon mellom undersøkelsene med hensyn til deltakernes fødeland. Andelen som er født i Norge er for eksempel 79 % i MoRo II og 98 % i HUNT II. For CONOR er andelen født i Norge 91,5 %, mens andelen i befolkningen som helhet er 93,5 % (Tabell 2). De som ikke har Norge som fødeland er i hovedsak født i de nordiske landene, Nord-Europa eller i land med historisk stor innvandring til Norge (Pakistan, Vietnam, Tyrkia, Sri Lanka, Iran, USA, Thailand, Filippinene). HUBRO og MoRo II har andelsmessig flere som er født i India, Pakistan og Øst-Asia, enn i de nordiske land/Nord-Europa. I Innvandrere-HUBRO deltok personer fra følgende fem land: Tyrkia (582 personer), Sri Lanka (1189 personer), Iran

(708 personer), Pakistan (546 personer) og Vietnam (657 personer). Vi har ikke sammenliknet fordelingen av fødeland for dem som ikke er født i Norge.

Røyking og alkohol

Midt i innsamlingsperioden for CONOR-data, i år 2000, var andelen som røykte daglig på landsbasis 32 % for begge kjønn. For menn gikk andelen som røyker daglig ned med omlag 1 % per år i hele innsamlingsperioden, men for kvinner var andelen daglig røykere rundt 32 % fram til år 2000 og ble deretter redusert tilsvarende som for menn. I CONOR rapporterte vel 30 % av begge kjønn at de var daglig røykere, men dette varierte sterkt mellom helseundersøkelsene. I Tromsø IV (1994-95) svarte om lag 37 % av mennene at de røykte daglig, mens bare 10,3 % av kvinnene i Innvandrers-HUBRO (2002) svarte det samme.

Forbruk av alkohol i Norge har vært forholdsvis stabilt de siste ti årene (5). I 2004 var andelen som ikke hadde drukket alkohol siste år 13 % for befolkningen (15 % for kvinner og 11 % for menn). Andel av CONOR-deltakerne som sier de ikke har drukket alkohol siste år eller er totalavholdende er 12,6 % (16,2 % for kvinner og 8,8 % for menn). Blant de i befolkningen som drikker alkohol, er frekvensen 1,1 ganger per uke, men her skiller ikke SIRUS mellom menn og kvinner. Tilsvarende CONOR-verdier er 1,3 ganger per uke for menn og 0,7 ganger per uke for kvinner, totalt 1,0 ganger per uke. Fra tabell 2 ser vi at HUBRO/Tromsø V er de undersøkelsene med høyest alkoholkonsum, mens Innvandrers-HUBRO er den undersøkelsen med lavest alkoholkonsum.

Hvor godt representerer CONOR den norske befolkningen?

For å undersøke forskjeller mellom de ulike helseundersøkelsene og landet for øvrig, har vi laget et avstandsmål. Vi har tatt utgangspunkt i andeler og har ikke vektet/justert for størrelsen

til helseundersøkelsene. For spørsmål med to kategorier, som daglig røyking, er avstandsmålet (d_i) differensen mellom andelen røykere i den aktuelle helseundersøkelse (p_i) og andelen i befolkningen ($d_i = p_i - p_{NORGE}$). For spørsmål med flere enn én kategori, som alder, beregnes avstandsmålet som summen av avstandene mellom andelene i de ulike kategoriene til helseundersøkelsene og tilsvarende andeler i landet for øvrig ($d_i = \sum_j (|p_{ij} - p_{NORGE,j}|)$). Den helseundersøkelsen som har lavest verdi for dette målet, ligger nærmest befolkningen. Beregning av avstandsmål for de ulike variablene er basert på tallene i tabellene 2 og 3. Vi viser ikke de beregnede verdiene. Siden CONORs deltakere er rekruttert i tidsrommet 1994 til 2003, og siden befolkningen har endret både røykevaner og bruk av alkohol i det samme tidsrommet, kan vi ikke uten videre sammenlikne CONOR med befolkningen. Våre sammenlikninger er derfor kun beskrivende. Statistisk sentralbyrås befolkningstall og røykedata er fra 2000, mens alkoholtallene fra SIRUS er fra 2004.

For alder har vi ekskludert både deltakere under 20 år og deltakere over 80 år. For denne variabelen er Tromsø IV og HUNT II nærmest befolkningen, mens HUSK, som inviterte 40-49 og 70-79 åringer, er lengst unna befolkningen (Tabell 2). CONOR samlet følger etter disse to førstnevnte helseundersøkelsene. Når det gjelder kjønns sammensetning er CONOR nærmere befolkningen enn noen av helseundersøkelsene, med et avvik fra befolkningen på ca ett prosentpoeng. For utdanning ligger CONOR svært nær gjennomsnittet i befolkningen og nærmere enn alle helseundersøkelsene. Tromsø V er den helseundersøkelsen der andelene deltakere født i og utenfor Norge ligger nærmest Norge. Også andelene i HUSK, OPPHED og Oslo 2 ligger nærmere landsgjennomsnittet enn CONOR, som ikke minst på grunn av Innvandrere-HUBRO, har en overrepresentasjon av personer født utenfor Norge.

For røyking er det HUSK som ligger nærmest befolkningen med CONOR på tredjeplass (Tabell 3). HUNT II og CONOR ligger nærmest landsgjennomsnittet når det gjelder andel som ikke har drukket alkohol siste år eller er totalavholdende.

Selv om CONOR består av flere tilnærmet tilfeldige utvalg av den voksne delen av Norges befolkning, er CONOR noe skjevfordelt på områder som alder og fødeland. Til tross for en relativt sett mindre andel deltakere blant de yngste og eldste var oppmøte høyest for disse gruppene. Det at CONOR ligger såpass nært landsgjennomsnittet for de andre variable vi har undersøkt, øker anvendbarheten til denne databasen.

For de helseutfallene som er registrert i CONOR, vil vi være forsiktig med beskrive befolkningen i Norge karakterisert ved insidens eller prevalens. Skjevheter i alder, fødeland og bosted (mange fylker er ikke representert) er relativt betydelig. Skal man ta høyde for disse skjevhetene, må man standardisere verdien for insidens/prevalens ved å vekte opp de gruppene som er underrepresentert og vekte ned de gruppene som er overrepresentert, se Rothman og Greenland, side 134 (6). For slike studier kan det være bedre å bruke andre registre der sykdomsutfallene finnes for den fullstendige befolkningen - som for eksempel Kreftregisteret, Norsk pasientregister eller Medisinsk fødselsregister.

Mulige epidemiologisk problemstillinger og design

CONOR er særlig nyttig for assosiasjonsstudier der vi vil studere om risiko for sykdom henger sammen med eksponering, enten ved å bruke hele databasen i kohort-, tverrsnitt- eller økologiske studier, eller deler av databasen i kasus-kontroll studier. For slike undersøkelser er det ikke så viktig om utvalget er representativt. Grunnen til dette er at vi nå ikke ønsker å beskrive den populasjonen utvalget er hentet fra (beskrivende epidemiologi), men

karakterisere en biologisk mekanisme mellom eksponering og sykdom (analytisk epidemiologi). I følge Rothman og Greenland (side 133) (6) og Rothman (side 20-21) (7), er det viktigere å velge design i analytiske epidemiologiske studier slik at man kan gjøre valide sammenlikninger mellom grupper, enn å velge design utfra at utvalgene skal være representative. Hvorvidt dette kravet er oppfylt for CONOR, må vurderes i hvert enkelt tilfelle.

Vi vil bruke et tenkt eksempel i diskusjonen av ulike epidemiologiske design. I dette eksempelet er målet å undersøke sammenhengen mellom lungekreft og radoneksponering: Man antar at om lag 10 % av lungekrefttilfellene skyldes radoneksponering (8), men disse anslagene kan være høyere i Norge, for eksempel på grunn av spesielle geologiske og klimatiske forhold. Vi tenker oss at vi har koblet CONOR med Kreftregisteret for å identifisere personer med lungekreft. Eksponering av radon kan hentes fra det digitaliserte kartverket Aktsomhetskartet (www.ngu.no/kart/arealis) som viser beregnede verdier for radioaktivitet fra radon i deler av Norge. Kartet er i hovedsak basert på noe over 35 000 radonmålinger i hus, luftmålinger av radioaktivitet og kjennskap til løsmasser i Norge (www.nrpa.no/radon). Koblingen tenker vi oss er gjort ved å bruke radonverdi knyttet til hver CONOR deltakers bostedsadresse (geocoding).

Økologiske studier

For å sammenlikne risiko for sykdom mellom områder kan man utføre økologiske studier for de områdene som CONOR dekker. For de samme områder kan man også produsere tematiske kart over sykdom (disease mapping). Man kan begrense seg til bestemte aldersgrupper slik at de ulike gruppene som skal sammenliknes i CONOR har omtrent samme aldersfordeling, men

man kan også her ta høyde for ulik aldersfordeling ved hjelp av statistiske metoder. Dette gjelder også prosjekter der man vil finne geografisk opphopning av sykdom (9).

For vårt eksempel kan et tematisk kart bidra til å visualisere sammenhengen mellom radoneksponering og lungekreft. CONOR kan også brukes til å få kunnskap om befolkningens eksponering av radon fordi det er naturlig å tenke seg at bebyggelsen er avhengig av de geologiske forholdene. Likevel vil man få best informasjon om befolkningens eksponering ved å koble Folkeregisteret og Aktsomhetskartet fordi en slik kobling vil inneholde hele landet.

Tverrsnittundersøkelser

CONOR kan også brukes som en *tverrsnittundersøkelse* der vi ser om deltakerne har sykdom ved ett gitt tidspunkt, og relatere forekomst av sykdom til ulike bakgrunnsfaktorer. Sjøgaard og medforfattere brukte CONOR på en slik måte for å undersøke sammenhengen mellom underarmsbrudd og urbaniseringsgrad – dvs hvor mange innbyggere det var i den kommunen deltakerne bodde (by, tettbygd eller spredt bebyggelse) (3).

Kohort studier

I *prospektive kohortstudier* er det mest naturlige å modellere tid til diagnose av lungekreft som en funksjon av radoneksponering og røyking, som jo er den viktigste faktoren for å utvikle lungekreft, samt legge inn andre forhold som alder, kjønn, yrke, fysisk aktivitet og alkoholbruk i en slik modell. Hvis forekomsten av lungekreft øker med radoneksponering anslått fra Aktsomhetskartet, er det en indikasjon på at radon faktisk er en årsak til lungekreft. Men det er mulig at individer som bor i områder med høy radoneksponering også røyker mer enn andre eller på annen måte er utsatt for faktorer som øker risikoen. Slike konfunderende

(sammenblandende) faktorer kan det justeres for med statistisk metodikk. Vi kan anta at vi vil fange opp den biologiske mekanismen mellom radonstråling og lungekreft fordi vi har deltakere fra ulike geologiske provinser. I hvert område har vi også deltakere som røyker daglig og deltakere som ikke røyker daglig.

Kasus-kontroll studier

Det kan tenkes at vi ikke er fornøyd med den eksponeringsvariabelen som vi presenterte over (gjennom bostedsadresse og Aktsomhetskartet) fordi boligtype, alder på bolig, ventilasjon og lufting vil være viktig for de radonverdiene vi måler i boligene. Det meste av denne informasjonen har vi ikke tilgang til i CONOR, og vi vil derfor utføre slike målinger i deltakernes boliger. Kostnader til utstyr og personer vil være store og vi vil derfor begrense oss til personer med lungekreft og et utvalg kontrollindivider som ikke har sykdommen. Det er flere varianter av slike kasus-kontroll studier og de adskiller seg ved hvordan kontrollene velges.

For problemstillingen i vårt eksempel kan en *nøstet kasus-kontroll* studie være aktuell (10). Her velger vi kontroller til et kasus blant alle deltakere i kohorten som ikke har utviklet lungekreft ved den alder og kanskje også det tidsrom kasus fikk sykdommen. Risiko for lungekreft øker sterkt med alder og vi unngår med denne utvalgsmetoden forskjeller i alder mellom kasus og kontroller.

Et alternativ til nøstede kasus-kontroll studier er *kasus-kohort* studier (10). Mekanismen for å velge kontroller i en kasus-kohort studie er å velge en gruppe deltakere i kohorten ved inklusjonstidspunktet uavhengig av om de senere blir kasus eller ikke. Denne kontrollgruppen kalles også en subkohort. Kasus registreres på vanlig måte blant personene i kohorten som

utvikler sykdom. Tradisjonelt har dette designet vært aktuelt for prosjekter der man registrerer mer enn ett utfall (sykdommer). Tanken har vært å samle informasjon, for eksempel biologiske markører, for en subkohort, og bruke denne subkohorten som kontroll for flere utfall. Det kan nevnes at Saarela og medarbeidere har utviklet statistiske metoder som kan gjøre oss i stand til å anvende også nøstede kasus-kontroll studier i prosjekter med flere utfall (11).

Når vi skal velge mellom de to designene som er nevnt over, er det også andre hensyn som må veies mot hverandre (12). For det første har tidsforskjellen mellom innsamlingstidspunkt og analysetidspunkt betydning. For kasus-kohort studier finner man verdier for eksponeringsdata og biologiske markører for subkohorten først, men man må ofte vente i flere år før sykdom inntreffer og data fra kasus kan analyseres. For nøstede kasus-kontroll studier derimot, analyserer man biologisk materiale for både kasus og kontroller på samme tidspunkt. For stabile markører slik som DNA har ikke dette noen betydning, men for mindre stabile agenter, som for eksempel fettstoffer i blodet, kan en nøstet kasus-kontroll studie passe bedre enn en kasus-kohort studie. For det andre har egenskaper til kohorten, som tid fra rekruttering og lengde av oppfølging, betydning. For kohorter som har rekruttert personer over tid (som CONOR), passer trolig nøstede kasus-kontroll studier bedre enn kasus-kohort studier, ifølge studier gjort av Borgan og Samuelsen (13). For å få et presist svar på dette må man gjennomføre en tilsvarende simuleringsstudie basert på CONOR, slik som Borgan og Samuelsen har gjort.

Stratifisering i kasus-kontroll studier

For sykdommer med sjeldne sammenhenger mellom eksponering og sykdom kan både nøstede kasus-kontroll og kasus-kohort studier styrkes ved stratifisering (14,15). Ideen er å

bruke en stratifiseringsvariabel, som er lett tilgjengelig for hele kohorten og korrelert med den aktuelle eksponeringsvariabelen, til å velge kontroller mer optimalt. Kohorten blir delt inn i strata ved hjelp av stratifiseringsvariabelen. For nøstede kasus-kontroll studier velges kontrollene avhengig av hvilket stratum kasus hører til: de velges både fra samme stratum som kasus og fra de andre strata (counter-matching). Dette vil maksimere variabiliteten til eksponeringsvariabelen og redusere usikkerheten til risikoanslaget for denne variabelen. For vårt tideligere eksempel med radon og lungekreft, kunne boligtype være en slik stratifiseringsvariabel fordi vi kan anta at boligtype er korrelert med radon i huset. For hvert kasus kunne vi valgt to kontroller: en som bor i enebolig og en som bor i leilighet.

Utvalgsberegning og styrke i gen-miljø eller gen-gen studier

Det finnes flere hel-genom assosiasjonsstudier (GWAS) publisert for sykdommer som Crohn, ulcerøs kolitt, diabetes, psoriasis, astma, allergi og osteoporose (16-20). Forskergrupper har funnet risikogener for sykdom (alleler) og funksjonaliteten til flere av disse genene. Risiko for sykdom for disse gen-variantene er riktignok svake (oddsforhold < 1.5 for mange av allelene), og det er fortsatt ubesvarte spørsmål knyttet til sykdommenes etiologi og samspillet mellom miljøvariable og genetikk. CONOR passer utmerket til slike studier fordi det er såpass stort antall deltakere og fordi vi har viktige eksponeringsvariable og biologiske markører, men spørsmålet er om antallet er stort nok?

For slike studier må vi inkludere et tilstrekkelig antall personer, slik at vi kan påvise en forventet sammenheng (forkaste nullhypotesen) med en så stor sannsynlighet som mulig (styrke). En slik antatt sammenheng kan være et risikoanslag mellom eksponering og sykdom, genetikk og sykdom eller samspill (interaksjon) miljø-gen og sykdom. For slike miljø-gen (og gen-gen) studier har Gauderman og Morrison utviklet et program (QUANTO) til å beregne

antall kasus og kontroller som må inkluderes (21). For studier der man vil se på triaden sykdom-gen-miljø må man ha en formening om følgende faktorer: prevalens av sykdom, andelen med aktuelle alleler i befolkningen, andelen som blir eksponert, spesifisering av genotypen (dominant / recessiv) og antatt risiko for sykdom og eksponering, sykdom og genetikk og interaksjon mellom disse. For noen sykdommer har vi tilstrekkelig antall kasus per dags dato, men for andre sykdomsutfall må man vente til antall kasus er tilstrekkelig stort. Vi vil ikke her gi eksempler på dette fordi det vil variere så mye fra problemstilling til problemstilling.

Diskusjon og konklusjon

Tanken bak CONOR var å få et stort nok antall deltakere til å studere årsaker til sjeldne sykdommer. De som er invitert til de forskjellige helseundersøkelsene, og som inngår i CONOR, har ulik sammensetning mht alder, kjønn og fødeland. Derfor blir fordelingen av disse variablene noe skjevfordelt i forhold til Norge som helhet. Andelen personer over 40 år (unntatt de mellom 50 og 59 år og de over 80 år) er forholdsmessig høyere enn i befolkningen. Det er flere kvinner enn menn i alle helseundersøkelsene unntatt Oslo II (der bare menn ble invitert) og Innvandrere-HUBRO. For røyking, bruk av alkohol og utdanning er tallene for CONOR og befolkningen relativt like.

CONOR er en samling av flere utvalg og blir således en heterogen kohort. Skjevhetene gjør at kohorten ikke passer for insidensstudier, men så lenge man kontrollerer (justerer) for de skjevhetene vi vet befinner seg i databasen, kan databasen brukes til økologiske, tverrsnitt og kohortstudier. For undersøkelser av sammenhenger mellom sykdom, miljø og genetikk representerer CONOR en rik database. For slike kostnadskrevenne prosjekter vil primært nøstede kasus-kontroll studier være aktuelle, men kasus-kohort studier kan også være

relevante for prosjekter som inkluderer flere sykdommer. En nærmere vurdering må gjøres i hvert enkelt prosjekt.

Referanser

1. Engeland A, Sjøgaard AJ. CONOR (COhort of NORway) - En oversikt over en unik forskningsdatabank. *Norsk Epidemiologi* 2003; 13: 73-7.
2. Magnus P, Andersen E, Holmen J et al. CONOR (Cohort NORway): historie, formål og potensiale. *Norsk Epidemiologi* 2003; 13: 78-82.
3. Sjøgaard AJ, Gustad TK, Bjertness E et al. Urban-rural differences in distal forearm fractures: Cohort Norway. *Osteoporos Int* 2007; 18: 1063-72.
4. Naess O, Sogaard AJ, Arnesen E et al. Cohort Profile: Cohort of Norway (CONOR). *Int J Epidemiol* 2007.
5. Horverak Ø, Bye EK. Det norske drikkemønsteret. En studie basert på intervjudata fra 1973-2004. SIRUSrapport 2. Statens institutt for rusmiddelforskning, 2007.
6. Rothman KJ, Greenland S. *Modern epidemiology*. 2. utg. Philadelphia: Lippincott Williams and Wilkins, 1998.
7. Rothman KJ. *Epidemiology An Introduction*. New York: Oxford University Press, 2002.
8. Copes R, Scott J. Radon exposure: can we make a difference? *CMAJ* 2007; 177: 1229-31.
9. Aamodt G, Samuelsen SO, Skrondal A. A simulation study of three methods for detecting disease clusters. *Int J Health Geogr* 2006; 5: 15.
10. Langholz B, Thomas DC. Nested case-control and case-cohort methods of sampling from a cohort: a critical comparison. *Am J Epidemiol* 1990; 131: 169-76.
11. Saarela O, Kulathinal S, Arjas E et al. Nested case-control data utilized for multiple outcomes: a likelihood approach and alternatives. *Stat Med* 2008; 27: 5991-6008.
12. Rundle AG, Vineis P, Ahsan H. Design options for molecular epidemiology research within cohort studies. *Cancer Epidemiol Biomarkers Prev* 2005; 14: 1899-907.
13. Borgan Ø, Samuelsen SO. A review of cohort sampling designs for Cox's regression model: Potentials in epidemiology. *Norsk Epidemiologi* 2003; 13: 239-48.
14. Langholz B, Borgan O. Counter-Matching - A Stratified Nested Case-Control Sampling Method. *Biometrika* 1995; 82: 69-79.
15. Borgan O, Langholz B, Samuelsen SO et al. Exposure stratified case-cohort designs. *Lifetime Data Analysis* 2000; 6: 39-58.
16. Aranzana MJ, Kim S, Zhao K et al. Genome-wide association mapping in *Arabidopsis* identifies previously known flowering time and pathogen resistance genes. *PLoS Genet* 2005; 1: e60.

17. Hakonarson H, Qu HQ, Bradfield JP et al. A novel susceptibility locus for type 1 diabetes on Chr12q13 identified by a genome-wide association study. *Diabetes* 2008; 57: 1143-6.
18. Liu Y, Helms C, Liao W et al. A genome-wide association study of psoriasis and psoriatic arthritis identifies new disease loci. *PLoS Genet* 2008; 4: e1000041.
19. Richards JB, Rivadeneira F, Inouye M et al. Bone mineral density, osteoporosis, and osteoporotic fractures: a genome-wide association study. *Lancet* 2008; 371: 1505-12.
20. Vercelli D. Advances in asthma and allergy genetics in 2007. *J Allergy Clin Immunol* 2008.
21. Gauderman WJ. Sample size requirements for matched case-control studies of gene-environment interaction. *Stat Med* 2002; 21: 35-50.

Tabell 1: Oversikt over de ulike helseundersøkelsene som inngår i CONOR, forkortelser, innsamlingsår og hvem som ble invitert

Navn	Forkortelse	Innsamlingsår	Inviterte – alder i år ¹⁾	Antall inviterte
Den fjerde Tromsøundersøkelsen	Tromsø IV	1994-1995	25 +	37 558
Helseundersøkelsen i Nord-Trøndelag II	HUNT II	1995-1997	20 +	94 196
Helseundersøkelsen i Hordaland	HUSK	1997-1999	40-44, 46-47, 70-72	38 587
Osloundersøkelsen II	Oslo II	2000	48-77	14 209 ²⁾
Helseundersøkelsen i Oslo	HUBRO	2000-2001	30, 31, 40, 45, 46, 59/ 60, 75/ 76	58 660 ³⁾
Helseundersøkelsen i Oppland og Hedemark	OPPHED	2000-2001	30, 40, 45, 60, 75	22 327
Den femte Tromsøundersøkelsen	Tromsø V	2001	30 +	10 353
Helseundersøkelsen i Oslo 2002 - Innvandrere	Innvandrere-HUBRO	2002	20-60	12 088 ⁴⁾
Helseundersøkelsene i Troms og Finnmark	TROFINN	2002	30-77	16 229 ⁵⁾
Morsjon på Romsås II	MoRo II	2003	34-70	5 535

1) Mer detaljert informasjon om utvalget som ble invitert til de enkelte helseundersøkelsene: <http://www.fhi.no/conor> - Cohort Norway (CONOR): Materials and methods

2) I tillegg kommer 2 515 menn som også tilhørte HUBRO og 50 menn som også tilhørte MoRo II. Disse ble bare invitert til HUBRO hhv Moro II.

3) Inkluderer 17 308 personer (31 og 46 år – tilleggskohort) som ikke ble purret. Deltakerne er ikke med i den ordinære HUBRO-filen.

4) Personer født i Pakistan, Tyrkia, Sri Lanka, Iran og Vietnam. Inkluderer 4 116 personer (20-30 år) som ikke ble purret.

5) Inkluderer 18 av 25 kommuner i Troms og 10 av 19 kommuner i Finnmark. De andre kommunene deltok i Tromsø V og SAMINOR. Den siste er ikke del av CONOR.

Tabell 2: Oversikt over fordeling av ulike sosio-demografiske nøkkelvariable for de ulike helseundersøkelsene og CONOR. Siste kolonne viser fordelingen til de samme variablene for befolkningen som helhet (for år 2000). Personer under 20 år er ekskludert.

		TROMSØ IV	HUNT II	HUSK	HUBRO	OPPHED	TROMSØ V	TROFINN	Oslo 2	MoRo II	Innvandrer- HUBRO	CONOR	NORGE
Alder (år)													
Menn	20-29	11,5	13,2	0,0	0,0	0,1	16,0	3,6	0,0	0,0	11,5	7,1	19,9
	30-39	25,0	17,8	0,2	28,0	14,0	46,3	9,1	0,0	14,6	44,2	16,5	22,3
	40-49	25,6	21,3	87,2	36,1	47,8	31,0	15,3	1,9	24,2	27,4	33,4	20,3
	50-59	17,3	17,4	0,0	9,9	0,3	2,0	31,3	8,5	40,0	16,1	13,3	17,6
	60-69	11,6	14,2	0,0	11,2	22,1	2,8	23,3	39,4	20,8	0,4	14,5	10,9
	70-79	7,4	11,9	12,6	14,8	15,7	2,0	17,1	50,2	0,5	0,0	15,2	9,0
	80+	1,7	3,8	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	-	-
	Totalt	47,8	47,2	45,7	44,3	45,6	44,0	46,3	100	44,2	52,0	48,7	49,8
Kvinner	20-29	12,2	13,9	0,0	0,0	0,3	19,0	5,8		0,0	23,6	8,4	18,1
	30-39	25,5	17,8	0,2	28,0	16,5	47,9	12,5		18,4	40,0	18,3	21,6
	40-49	23,8	20,3	86,4	37,3	48,2	22,3	19,2		24,3	21,7	36,3	19,7
	50-59	15,6	16,4	0,0	8,6	0,3	1,6	33,2		41,2	13,4	12,6	18,1
	60-69	11,4	13,5	0,0	10,1	20,5	4,5	24,8		15,9	0,4	11,6	11,7
	70-79	8,7	12,5	13,4	16,0	14,2	4,5	20,4		0,2	0,0	12,8	10,8
	80+	2,9	5,2	0,0	0,0	0,0	0,2	0,0		0,0	0,0	-	-
	Totalt	52,2	52,8	54,3	55,7	54,4	56,0	53,7		55,8	48,0	51,3	50,2
Utdanning													
Menn	Antall år (gjennomsnitt)	11,7	11,0	12,1	14,1	11,7	14,0	10,4	12,3	12,3	12,5	11,8	11,9
	Standard avvik	3,0	2,6	3,1	4,0	3,5	4,3	3,6	3,8	3,8	4,3	3,4	2,2
Kvinner	Antall år (gjennomsnitt)	11,5	10,9	11,8	13,7	12,0	14,0	10,6	-	11,9	10,9	11,6	11,6
	Standard avvik	3,0	2,6	3,0	4,1	3,5	4,4	3,7	-	3,6	4,7	3,3	2,1
Fødeland	Født i Norge (%)	95,6	97,6	95,3	82,2	95,7	92,1	96,4	95,4	78,5	0,0	91,5	93,5

Tabell 3: Oversikt over fordeling av nøkkelvariablene daglig røyking og bruk av alkohol for de ulike helseundersøkelsene og CONOR. Siste kolonne viser fordelingen til de samme variablene for befolkningen som helhet (røyking år 2000 og alkohol år 2004). Personer under 20 år er ekskludert.

		TROMSØ IV	HUNT II	HUSK	HUBRO	OPPHED	TROMSØ V	TROFINN	Oslo 2	MoRo II	Innvandrer- HUBRO	CONOR	NORGE
Røyking													
Menn	Røyker daglig (%)	37,4	28,8	33,1	25,9	29,1	35,1	34,3	20,3	28,8	34,6	30,1	32
Kvinner	Røyker daglig (%)	36,4	29,0	32,6	26,4	30,8	34,6	35,8	-	32,2	10,3	30,6	32
Alkoholkonsum													
Menn													
	Totalavhold/Ikke drukket siste år (%)	9,0	8,3	8,4	7,9	7,6	4,1	9,1	6,9	12,9	30,8	8,8	11
	Antall ganger med alkohol per uke (gjennomsnitt)	1,19	0,92	1,39	2,06	1,11	1,33	1,02	2,17	1,66	0,82	1,30	1,13
Kvinner													
	Totalavhold/Ikke drukket siste år (%)	16,8	16,6	12,9	11,1	13,9	10,3	19,9	-	15,1	71,9	16,2	15
	Antall ganger med alkohol per uke (gjennomsnitt)	0,75	0,55	0,87	1,48	0,70	0,74	0,59	-	1,06	0,11	0,79	1,13