TUTKIMUKSIA FORSKNINGSRAPPORTER RESEARCH REPORTS 239

TERVEYS 2000 -TUTKIMUS

AIKUISVÄESTÖN HAASTATTELUAINEISTON TILASTOLLINEN LAATU

Otanta-asetelma, tiedonkeruu, vastauskato ja estimointi- ja analyysiasetelma

Johanna Laiho ja Tarja Nieminen (toim.)

Tilastokeskus
Statistikcentralen
Statistics Finland

TUTKIMUKSIA FORSKNINGSRAPPORTER RESEARCH REPORTS 239

TERVEYS 2000 -TUTKIMUS

AIKUISVÄESTÖN HAASTATTELUAINEISTON TILASTOLLINEN LAATU

Otanta-asetelma, tiedonkeruu, vastauskato ja estimointi- ja analyysiasetelma

Johanna Laiho ja Tarja Nieminen (toim.)

Tilastokeskus
Statistikcentralen
Statistics Finland

Toimittajat
Johanna Laiho
Tarja Nieminen

Kannen suunnittelu
Maija Sohlman

Taitto
Outi Stenbäck

ISSN 0355–2071
ISBN 952–467–267–7

Yliopistopaino

Helsinki 2004

TERVEYS 2000 -TUTKIMUS

AIKUISVÄESTÖN HAASTATTELU-AINEISTON TILASTOLLINEN LAATU

Otanta-asetelma, tiedonkeruu, vastauskato ja estimointi- ja analyysiasetelma

Johanna Laiho ja Tarja Nieminen (toim.)

Alkusanat

Terveys 2000 -tutkimus on Kansanterveyslaitoksen johdolla toteutettu suurtutkimus suomalaisten terveydentilasta. Sen tiedot kerättiin vuonna 2000. Tiedot koostuvat erilaisista aineistoista. Tilastokeskus vastasi terveyshaastatteluaineiston tiedonkeruusta. Lisäksi Tilastokeskus osallistui tietojärjestelmän suunnitteluun, otanta-asetelman kehittämiseen ja tilastollisten menetelmien neuvontaan. Tämän työn suunnittelusta ja toteuttamisesta vastasi otanta, tiedonkeruu, tietojenkäsittely ja analyysi -työryhmä, johon kuuluivat Risto Lehtonen (pj), Kari Djerf, Pirjo Hyytiäinen, Tommi Härkänen, Paul Knekt, Kari Kuulasmaa, Vesa Kuusela, Johanna Laiho, Marjo Laine, Paula Lamberg, Jouni Maatela, Tuija Martelin, Erkki Nenonen, Mikko Nenonen, Tarja Nieminen, Mikko Nissinen, Timo Peltomaa, Harri Rissanen, Pentti Salmela, Matti Sarjakoski, Eero Tanskanen, Vesa Tanskanen, Tuula Tiainen, Kai Vikki ja Esa Virtala.

Lisäksi työn toteuttamiseen osallistui Kansanterveyslaitoksesta, Tilastokeskuksesta ja muista yhteistyöorganisaatioista suuri joukko eri aihealueiden asiantuntijoita, joille lämmin kiitos.

Tässä raportissa kuvataan Tilastokeskuksen tuottamien aineistojen laatua. Raportin kirjoittamiseen ovat osallistuneet Tilastokeskuksesta erikoistutkija Johanna Laiho, yliaktuaari Tarja Nieminen, kehittämispäällikkö Kari Djerf ja erikoistutkija Vesa Kuusela sekä professori Risto Lehtonen Jyväskylän yliopistosta ja tutkija Tommi Härkänen Kansanterveyslaitoksesta. Raportin taitosta vastasi Outi Stenbäck ja oikoluvusta Jaana Huhta.

Jussi Simpura
tilastojohtaja
Elinolot-yksikkö

Sisältö

Johdanto

Johanna Laiho

Terveys 2000 -tutkimuksessa on kerätty väestön terveystietoja haastatteluiden, itse täytettävien kyselyiden ja kliinisten tutkimusten avulla. Tutkimushankkeen pääkoordinaattori on Kansanterveyslaitos (KTL), ja hanke on toteutettu laajan tutkimusyhteistyön puitteissa. KTL:n lisäksi tutkimuksen rahoittajina ja toteuttajina toimivat myös Eläketurvakeskus, Kansaneläkelaitos, Kuntien eläkevakuutus, Stakes, Suomen Hammaslääkäriliitto, Suomen Hammaslääkäriseura, Tilastokeskus, Työsuojelurahasto, Työterveyslaitos, UKK-instituutti ja Valtion työsuojelurahasto. Myös monet yritykset ovat osallistuneet tutkimuksen toteutukseen laitetoimittajina. Aromaa ja Koskinen (2002) ovat kuvanneet tätä raporttia tarkemmin Terveys 2000 -tutkimuksen projektiorganisaatiota, taustaa ja tutkimusasetelmaa. Lisäksi KTL:n valmisteilla oleva menetelmäraportti keskittyy tutkimusasetelman toteutumisen kuvaukseen ja arviointiin.

Terveys 2000 -tutkimus sisältää useita osia: 30 vuotta täyttäneen aikuisväestön perusteellisen terveystutkimuksen, nuorten aikuisten haastattelututkimuksen (18–29-vuotiaat) sekä Mini-Suomi-tutkimuksen seurantahankkeen. Tilastokeskus on osallistunut kahden ensiksi mainitun osion tiedonkeruuseen. Tilastokeskuksen rooli on painottunut tutkimuksen kenttätyön ja tietojärjestelmän suunnitteluun, terveyshaastatteluiden toteuttamiseen, otanta-asetelman suunnitteluun, tutkimusaineistojen painokertoimien muodostamiseen ja tilastollisten menetelmien neuvontaan.

Tässä raportissa arvioidaan 30 vuotta täyttäneen aikuisväestön terveyshaastatteluosion onnistumista. Otantatutkimuksen kokonaislaatuun vaikuttavat useat osatekijät. Tutkimusaineiston hyvä tilastollinen laatu edellyttää muun muassa virheetöntä otantakehikkoa, kaikkien kohdeperusjoukon väestöryhmien korkeata vastausosuutta, yksiselitteisesti ymmärrettyä kyselylomaketta, hyvin koulutettua ja ammattitaitoisista haastattelijakuntaa, haastattelijavaikutuksen määrätietoista minimointia ja mittausvirheiden huolellista ennaltaehkäisyä (Laiho, 2002a).

Laatuselvityksen ensimmäisessä luvussa kerrotaan lyhyesti Terveys 2000 -tutkimuksen taustasta. Toisessa luvussa kuvataan haastattelujen toteutus ja tiedonkeruu. Kyseisessä luvussa käsitellään myös sähköistä tiedonkeruuta ja -siirtoa. Otanta-asetelma esitetään kolmannessa luvussa ja katoanalyysi neljännessä luvussa. Viidennessä luvussa kuvataan terveyshaastattelutietojen painokertoimien muodostaminen ja tarkastellaan asetelmapohjaista estimointia. Laatuselvityksen kuudennessa luvussa esitellään monimutkaisen otanta-asetelman asettamat vaatimukset aineiston käytölle sekä annetaan ohjeita ja esimerkkejä tutkimusaineiston käytöstä ja analysoinnista. Luvun yhteenvedossa verrataan eri analyysimenetelmiä ja perustellaan annetut suositukset Terveys 2000 -tutkimusaineiston tilastolliselle analysoinnille.

1 Terveys 2000 -tutkimuksen tausta

Tarja Nieminen

Väestön terveyttä ja sen kehitystä seuraamalla saadaan tietoja, joiden avulla voidaan edistää kansanterveyttä, kehittää terveyspolitiikkaa sekä suunnitella terveyspalveluja ja sosiaaliturvaa. Terveys 2000 -tutkimuksen tarkoituksena on tuottaa ajantasaista ja tarkkaa tietoa Suomessa vakituisesti asuvan työikäisen ja iäkkään väestön terveydestä ja toimintakyvystä. Keskeisinä tutkimuskohteina ovat tärkeimmät kansansairaudet ja mielenterveys sekä niihin liittyvän hoidon, kuntoutuksen ja avun tarve. Lisäksi tuotetaan tietoa terveyteen vaikuttavista tekijöistä. Terveys 2000 -tutkimuksen aineistoa tullaan käyttämään poikkileikkausaineistona sekä ajallisena vertailuaineistona. Vertaamalla Terveys 2000 -tutkimuksen ja Mini-Suomi-tutkimuksen tuloksia voidaan tutkia terveyden ja toimintakyvyn kehittymistä ja laatia arvioita tulevaisuuden kehitysnäkymistä. (Aromaa ja Koskinen, 2002).

Kelan tekemä Mini-Suomi-tutkimus, johon monessa yhteydessä viitataan, toteutettiin 20 vuotta ennen Terveys 2000 -tutkimusta. Tutkimuksen kohteena oli 30 vuotta täyttänyt väestö. Otokseen poimittiin 8 000 henkilöä 40 alueelta eri puolilta Suomea. Kenttätyö alkoi vuonna 1978 haastatteluilla ja terveystarkastuksilla ja päättyi syventäviin tutkimuksiin vuonna 1981. Terveyshaastattelu tehtiin haastateltavan kotona tai laitoksessa, ja haastattelijoina toimivat paikalliset terveydenhoitajat ja sairaanhoitajat. Kaikki haastatellut kutsuttiin Kelan autoklinikassa tehtyyn terveystarkastukseen. Autoklinikka oli liikkuva tutkimusasema, johon kuului useita kulkuneuvoja ja tutkimushenkilökuntaa. Terveystarkastuksen perusteella seulotut kutsuttiin vielä jälkitutkimukseen. Lisäksi osa 30–64-vuotiaista kutsuttiin syventäviin tutkimuksiin Kelan kuntoutustutkimuskeskukseen Turussa. Haastatteluun osallistui 96 % ja terveystarkastuksen perustutkimukseen 90 % otokseen kuuluneista 8 000 henkilöstä. (Aromaa et al. 1989).

Terveys 2000 -tutkimuksen tiedot kerättiin usealla eri menetelmällä: haastatteluiden, itse täytettävien kyselylomakkeiden, verenpaineen kotimittauksen ja kliinisen terveystarkastuksen avulla. Osalle terveystarkastukseen osallistuneista tehtiin vielä täydentäviä kliinisiä tutkimuksia yliopistosairaaloiden alueilla eräiltä erikoisaloilta. Lisäksi tutkimuksessa käytetään joitakin keskeisiä rekisteritietoja.

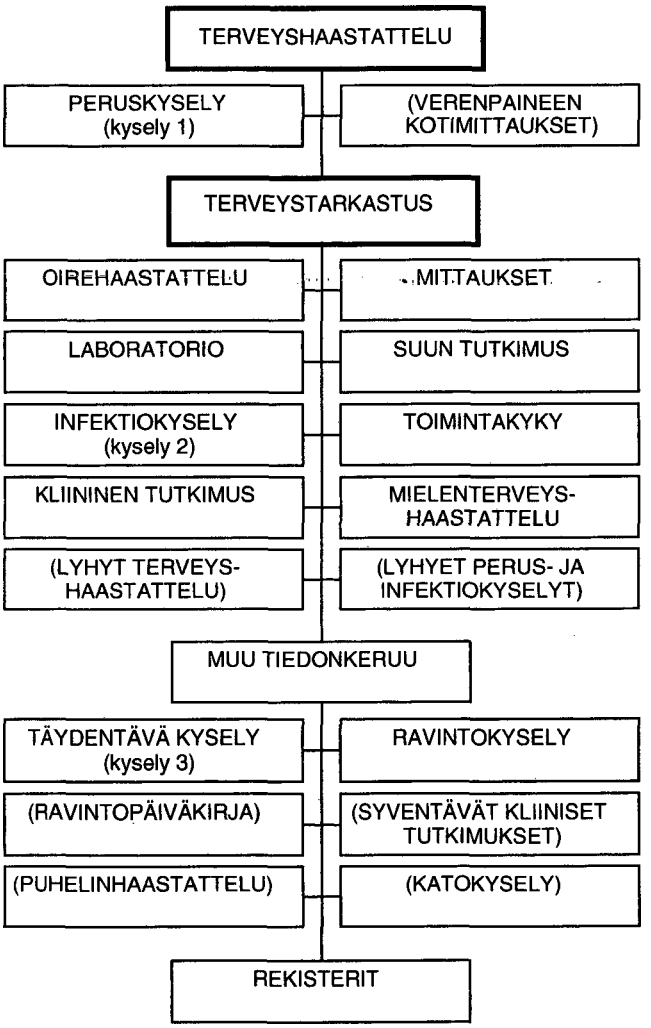
Tutkimus sisälsi kolme osaa:

1. 30 vuotta täyttäneiden tutkimus
2. 18–29-vuotiaiden nuorten aikuisten tutkimus ja
3. Mini-Suomi-tutkimukseen osallistuneiden seurantatutkimus.

Seurantatutkimukseen osallistuneiden sekä nuorten aikuisten tutkimukset olivat hieman suppeampia kuin 30 vuotta täyttäneiden tutkimus. Nuorille aikuisille ei tehty ollenkaan terveystarkastusta. Tilastokeskus teki osien 1–2 haastattelut. Kansanterveyslaitos (KTL) toteutti 30 vuotta täyttäneiden terveystarkastukset

sekä Mini-Suomi-tutkimukseen osallistuneiden uusintatutkimuksen kokonaisuudessaan.

Kuvio 1.1.
30 vuotta täyttäneiden tutkimuksen tiedonkeruuvaiheet Terveys 2000 -tutkimuksessa (suluissa olevat osiot kohdistettiin osalle)



Kuvio 1.1. kuvaa tiedonkeruuta 30 vuotta täyttäneiden tutkimuksessa. Useimmat kuviossa mainitut osat oli tarkoitettu kaikille otokseen kuuluneille (N=8 028). Suluissa olevat tiedonkeruuosiot kohdistettiin vain osalle otokseen kuuluneista.
Tiedonkeruu alkoi tutkittavien kotona tehdyllä haastattelulla, jonka yhteydessä haastattelijat myös varasivat haastatelluille ajan terveystarkastukseen. Lisäksi

vastaajilta pyydettiin kirjallinen lupa eräiden keskeisten rekisteritietojen yhdistämiseen heidän tutkimustietoihinsa. Haastattelun jälkeen haastatelluille jätettiin täytettäväksi peruskysely (kysely 1) sekä osalle verenpainemittari ja verenpaineen seurantalomake (ks. luku 2).

Kansanterveyslaitoksen järjestämään terveystarkastukseen kuului oirehaastattelu, useita erilaisia mittauksia (esimerkiksi pituus, paino ja EKG), laboratoriotutkimuksia, suun ja hampaiden tutkimus ja röntgen, toimintakykytutkimus, lääkärin tekemä kliininen tutkimus sekä mielenterveyshaastattelu. Taukojen aikana tutkittavat täyttivät infektiokyselyn (kysely 2). Mikäli tutkittava henkilö ei ollut ennen terveystarkastusta osallistunut Tilastokeskuksen haastattelijoiden tekemään haastatteluun, hänet haastateltiin terveystarkastuksen yhteydessä. Tämä haastattelu oli jonkin verran varsinaista kotihaastattelua lyhyempi.

Terveystarkastuksen päätteeksi tutkituille jaettiin täydentävä kysely (kysely 3) ja ravintokysely sekä osalle ravintopäiväkirja, jotka pyydettiin täyttämään kotona ja palauttamaan Kansanterveyslaitokseen postitse valmiissa palautuskuoressa. Osa tutkituista kutsuttiin vielä jälkeenpäin syventäviin kliinisiin tutkimuksiin. Lisäksi UKK-instituutti kutsui Tampereen seudulla asuvat tutkitut toimintakykytutkimukseen. Terveystarkastuksesta on kerrottu tarkemmin Terveys 2000 -tutkimuksen perustulosraportissa (Aromaa ja Koskinen, 2002) sekä parhaillaan valmisteltavassa KTL:n menetelmäraportissa Terveys 2000 -tutkimuksen toteutus, aineisto ja menetelmät.

Jos tutkimukseen osallistuva henkilö ei pystynyt esimerkiksi sairauden takia osallistumaan pitkään terveystarkastukseen ja haastatteluun, hänelle tehtiin lyhennetty terveystarkastus ja haastattelu kotona. Samalla hänelle jätettiin lyhyet perus- ja infektiokyselyt (lyhyet kyselyt 1 ja 2).

Niille, jotka eivät osallistuneet edellä mainittuihin tutkimusosioihin, KTL pyrki tekemään lyhyen puhelinhaastattelun, joka sisälsi eräitä keskeisiä terveysmittareita. Jos puhelinhaastatteluakaan ei saatu, KTL lähetti katokyselyn. Näin suurimmalta osalta otokseen kuuluneista saatiin ainakin joitakin keskeisimpiä terveystietoja.

Ennen varsinaista tutkimusta tehtiin kaksi pilottitutkimusta haastattelukysymysten ja kenttävaiheen prosessin testaamiseksi. Ensimmäinen pilotti toteutettiin tammi–helmikuussa ja toinen huhti–toukokuussa 2000. Pilottien tulosten ja pilottiin osallistuneiden Tilastokeskuksen haastattelijoiden arvioiden perusteella tehtiin tarkistuksia suunnitelmiin, ohjeisiin sekä haastattelulomakkeeseen. Varsinainen Terveys 2000 -tutkimuksen kenttätyö alkoi Tilastokeskuksen haastattelijoiden tekemillä terveyshaastatteluilla, joista kerrotaan tarkemmin luvussa 2.

2 Terveyshaastattelu

Tarja Nieminen ja Vesa Kuusela

2.1 Terveyshaastattelun toteutus

Tilastokeskuksella on noin 160 vakinaisen, tehtäväänsä koulutetun haastattelijan muodostama haastattelijaorganisaatio. Haastattelijoita on koko maan alueella väestömäärän suhteessa. Kullekin haastattelijalle on määritelty oma pysyvä haastattelualueensa. Terveys 2000 -tutkimuksessa osa haastattelijoista työskenteli poikkeuksellisesti myös muilla alueilla, jotta alueittain porrastettua terveystarkastusaikataulua pystyttiin noudattamaan. Haastatteluiden tekemiseen osallistui yhteensä 158 haastattelijaa. Suurimmalla osalla heistä oli aikaisempaa kokemusta terveystutkimuksista. Esimerkiksi lähes 60 % heistä osallistui Stakesin ja Kelan Terveydenhuollon väestötutkimukseen 1995/96. Terveys 2000 -tutkimus oli kuitenkin tätä laajempi ja monipuolisempi kokonaisuus ja siksi myös vaativampi.

Terveys 2000 -tutkimusta varten haastattelijat saivat päivän mittaisen koulutuksen. Koulutus toteutettiin kolmessa noin 50 haastattelijan ryhmässä. Kouluttajina toimivat eri aihealueiden asiantuntijat Kansanterveyslaitoksesta ja eräistä muista tutkimukseen osallistuneista organisaatioista. Koulutuksen lisäksi haastattelijoille annettiin kirjalliset työohjeet, joissa kerrottiin tutkimuksen taustasta, tarkoituksesta ja sisällöstä sekä annettiin käytännön toimintaohjeita ja sisältöön liittyviä ohjeita. Joitakin kysymyskohtaisia ohjeita oli merkitty myös haastattelulomakkeelle kysymysten yhteyteen. Haastattelijoille annettiin lisäksi lista yhteyshenkilöistä, joilta saattoi kysyä neuvoja kenttätyön aikana.

Haastattelut tehtiin 15.8.2000–28.2.2001. Kenttätyöaika vaihteli paikkakunnittain terveystarkastusaikataulun mukaan. Haastattelut oli porrastettu eri paikkakunnilla siten, että ne ehdittiin tehdä ennen terveystarkastusten alkamista tai suuremmissa kaupungeissa ennen niiden päättymistä. Haastattelut käynnistyivät terveyskeskuspiireissä yleensä noin kaksi kuukautta ennen terveystarkastusten alkamista, joskin tästä säännöstä poikettiin tarvittaessa. Kussakin terveyskeskuspiirissä haastatteluihin oli varattu aikaa yhdestä kolmeen kuukauteen. Tänä aikana tavoittamatta jääneitä henkilöitä yritettiin tavoitella helmikuun 2001 loppuun asti.

Haastatteluprosessi eteni seuraavasti:

1) Haastattelija lähetti tutkimuksesta kertovan kirjeen ja esitteen otokseen kuuluvalle henkilölle.
2) Haastattelija soitti kirjeen vastaanottajalle sopiakseen haastatteluajasta. Jos puhelinnumeroa ei löytynyt, kirjeeseen merkittiin ehdotus haastatteluajasta sekä soittopyyntö, jos aika ei sovi.
3) Käyntihaastattelu, jonka yhteydessä:

- ajan varaaminen terveystarkastukseen sekä ajanvaraus- ja osoitetietojen jättäminen paperilla haastatellulle,
- tutkimuksesta kertovan tiedotteen ja suostumuslomakkeen antaminen, allekirjoituksen ottaminen suostumuslomakkeeseen,
- verenpainemittarin, ohjeiden ja mittausten seurantalomakkeen jättäminen, mittarin käytön opetus sekä
- peruskyselyn (kyselyn 1) ja palautuskuoren jättäminen haastatellulle.

Haastattelut tehtiin pääsääntöisesti haastateltavan kotona (lähes 90 %). Jos tämä ei onnistunut, haastattelut pyrittiin tekemään jossakin muussa haastateltavalle sopivassa paikassa, esimerkiksi Tilastokeskuksessa tai paikallisella terveysasemalla. Myös laitoksissa olevat henkilöt haastateltiin. Laitoksissa haastatteluista tehtiin 9 %. Vain poikkeustapauksissa haastattelu tehtiin puhelimitse.

Haastattelulomake (ks. www.ktl.fi/terveys2000) sisälsi kymmenen osiota:
- taustatiedot,
- terveydentila ja sairaudet,
- vanhempia ja sisaruksia koskevat kysymykset,
- terveyspalvelut,
- suun terveys,
- elintavat,
- elinympäristö,
- toimintakyky,
- työ ja työkyky sekä
- kuntoutus.

Lisäksi lomakkeessa oli haastattelijan täytettäväksi tarkoitettuja kysymyksiä mahdollisista sijaisvastaajista ja annettujen tietojen luotettavuudesta.

Haastattelun kesto oli keskimäärin 95 minuuttia. Haastattelulomake oli tehty suomeksi ja ruotsiksi. Toinen näistä oli myös yleisimmin vastaajien äidinkieli (98,2 %). Muista vastaajien äidinkielistä yleisimpiä olivat saame, venäjä ja viro. Lisäksi haastateltiin yksittäisiä muunkielisiä henkilöitä. Jos haastateltava ei ymmärtänyt riittävästi suomea tai ruotsia, haastattelu tehtiin tulkin avulla. Haastattelun tekemisen edellytyksenä oli, että haastattelusta pystyttiin sopimaan jollakin yhteisellä kielellä. Tulkkeina käytettiin tapauksesta riippuen perheenjäsentä, sukulaista, muuta tuttua tai ammattitulkkia. Tulkkia vaatineiden haastatteluiden määrä oli kuitenkin pieni.

Sijaisvastaajia käytettiin haastattelutilanteessa poikkeustapauksissa, esimerkiksi jos varsinainen haastateltava oli liian huonossa kunnossa vastatakseen itse. Sijaisvastaajalla tarkoitetaan henkilöä, joka tuntee tutkittavan ja voi tarvittaessa antaa ainakin osan haastattelutiedoista hänen puolestaan. Sijaisvastaajan sallitaan vastata ainoastaan tosiasiakysymyksiin, mutta ei mielipidekysymyksiin. Sijaisvastaaja voi myös auttaa haastateltavaa ja vastata yhdessä hänen kanssaan esimerkiksi antamalla tietoja, joita haastateltava itse ei muista. 197 haastateltavaa (2,8 %) vastasi jonkun lähiomaisen avustamana.

Osassa haastatteluista (2 %) haastateltava ei kyennyt ollenkaan vastaamaan itse, jolloin jouduttiin käyttämään pelkästään sijaisvastaajaa. Sijaisvastaajan käyttäminen yleistyi kohdehenkilön iän mukana, mutta vasta tutkittavien iän ylittäessä 89 vuotta sijaisvastaajat vastasivat hieman useammin kuin kohdehen-

kilöt itse. Sijaisvastaajana toimi yleensä puoliso tai lapset. Myös hoitohenkilökunta, sukulaiset tai ystävät antoivat – tutkittavan henkilön tai omaisten luvalla – tärkeimpiä terveyteen liittyviä haastattelutietoja.

Haastattelijoita pyydettiin jokaisen haastattelun jälkeen arvioimaan haastattelussa saatujen vastausten luotettavuutta, ja merkitsemään arvionsa haastattelulomakkeen loppuun. Heidän arvionsa mukaan suurin osa vastauksista oli luotettavia (93 %) tai osittain luotettavia (lähes 7 %). Luotettavuusongelmista kaksi kolmasosaa johtui muistivaikeuksista. Tulos ei ole yllättävä, koska haastattelu sisälsi muistia vaativia kysymyksiä sairauksista, työhistoriasta, vanhemmista ja lapsuuden oloista. Vain muutaman henkilön vastaukset arvioitiin epäluotettaviksi. Noin 9 prosentilla vastaajista oli vaikeuksia ymmärtää yksittäisiä kysymyksiä.

Ajanvaraus terveystarkastukseen

Haastatelluille varattiin haastattelun yhteydessä aika Kansanterveyslaitoksen järjestämään terveystarkastukseen. Jokaisella haastattelijalla oli terveystarkastusajoista yksilöllinen terveyskeskuspiirikohtainen varauslista, jonka avulla kullekin tutkittavalle varattiin oma henkilökohtainen aika, joka merkittiin haastattelulomakkeelle. KTL sai ajanvaraustiedot haastattelutietojen mukana ja välitti ne edelleen kenttäryhmille eri puolille maata.

Haastattelijat merkitsivät Kansanterveyslaitosta ja kenttäryhmiä varten lisätietoja terveystarkastusta varten, esimerkiksi jos haastateltava tarvitsi tulkkia tai liikkui pyörätuolilla. Alkuun tämän järjestelmän toiminnassa oli puutteita, koska oikeiden tietojen poimiminen ja toimittaminen eteenpäin kangerteli, mutta ongelma korjaantui tutkimuksen edetessä.

Noin 10 % käytettävissä olevista terveystarkastusajoista oli jätetty KTL:n ajanvarauskeskukseen mahdollisia ajan vaihtoja varten ja siltä varalta, että haastattelija tarvitsisi haastatellulle sellaisen ajan, jota hänen henkilökohtaisella ajanvarauslistallaan ei ollut. Jos ajanvarauslistalta ei löytynyt haastatellulle sopivaa aikaa, haastattelija otti yhteyttä puhelimitse KTL:n ajanvarauskeskukseen ja sai sieltä uuden ajan.

Jos haastattelu oli sovittu alle kymmenen päivää ennen terveystarkastusten päättymistä ao. paikkakunnalla, haastattelija ei saanut enää käyttää omaa ajanvarauslistaansa vaan hänen oli soitettava KTL:n ajanvaraukseen. Tämä käytäntö otettiin käyttöön myöhemmin haastatteluiden ollessa käynnissä, kun havaittiin ongelmia ajanvaraustietojen välittymisessä ajoissa terveystarkastusryhmiin haastattelun ja terveystarkastuksen ollessa hyvin lähekkäin toisiaan.

Osa otokseen kuuluvista henkilöistä osallistui Tilastokeskuksen haastatteluun, mutta ei halunnut mennä terveystarkastukseen tai päinvastoin. Jos henkilö ei osallistunut haastatteluun, mutta halusi mennä terveystarkastukseen, haastattelija varasi hänelle terveystarkastusajan. Suurin osa tutkittavista osallistui kuitenkin molempiin osioihin. Vastaavasti kieltäytyminen koski usein molempia osioita.

Laitoksissa oleville pyrittiin järjestämään terveystarkastus kyseisessä laitoksessa. Jos tutkittava asui kotona, mutta oli liian huonokuntoinen tullakseen terveystarkastuspaikkaan, terveystarkastus tehtiin kotona niiltä osin kuin se oli mahdollista.

Haastattelun yhteydessä terveystarkastusaika saatiin varattua 93 %:lle haastatelluista. Muille aika jäi varaamatta siksi, että tarjotut haastatteluajat eivät sopineet (21 %), haastateltu ei päässyt muusta syystä tulemaan (36 %) tai haastateltava ei halunnut (40 %) ollenkaan osallistua terveystarkastukseen.

Peruskysely (kysely 1)

Peruskyselyssä oli kysymyksiä elintavoista, työstä, elinympäristöstä, terveydestä ja hyvinvoinnista. Kyselylomake jätettiin haastatelluille haastattelun jälkeen kotiin täytettäväksi, ja se pyydettiin palauttamaan terveystarkastukseen. Lähes kaikki haastatellut ottivat kyselyn täytettäväksi. Haastatteluun vastanneista 196 (2,8 %) ei halunnut kyselyä.

Kysely jätettiin myös sellaisille henkilöille, jotka eivät tulleet terveystarkastukseen. Nämä henkilöt palauttivat täyttämänsä kyselyn palautuskuoressa postitse Kansanterveyslaitokseen. Jos varsinainen terveyshaastattelu tehtiin puhelimitse, haastattelija toimitti kyselyn haastatellulle, jota pyydettiin palauttamaan lomake täytettynä terveystarkastuksessa tai postitse palautuskuoressa. Kotiterveystarkastuksen yhteydessä täytettiin lyhennetyt perus- ja infektiokyselyt, joiden avulla saatiin tärkeitä tietoja, mutta jotka eivät olleet liian raskaat sairaille vastaajille.

Tarvittaessa joku läheinen henkilö tai terveystarkastuksen henkilökunta auttoi kyselyn täyttämisessä, jos haastateltavalla oli jokin vastaamista vaikeuttava vamma, esimerkiksi heikko näkökyky. Kyselyn vastaanottaminen oli yleistä myös tapauksissa, joissa haastateltava vastasi toisen henkilön avustamana. Lisäksi yli puolet sijaisvastaajista otti kyselyn täytettäväksi.

Muut tutkimukseen kuuluneet kyselylomakkeet annettiin terveystarkastuksessa. Aluksi hoitaja teki lyhyen oirehaastattelun. Infektiokysely (kysely 2) täytettiin odotushuoneessa eri tutkimusjaksojen välissä. Täydentävä kysely (kysely 3), ravintokysely ja ravintopäiväkirja annettiin terveystarkastuksen lopuksi kotiin täytettäviksi, ja ne pyydettiin palauttamaan postitse. Terveystarkastuksen vaiheista on kerrottu tarkemmin perustulosraportissa (Aromaa ja Koskinen, 2002).

Kotona tehtävä verenpaineen mittaus

Osalle haastatelluista jätettiin verenpainemittari kotona tapahtuvaa verenpaineen mittausta varten. Mittarin käyttämiseen koulutetut haastattelijat näyttivät, miten mittaus tehdään. Lisäksi haastatelluille annettiin kirjalliset ohjeet. Mittarin saaneita pyydettiin mittaamaan verenpainettaan viikon ajan aamulla ja illalla sekä kirjaamaan tulokset paperille. Verenpainemittareista ja mittauksen onnistumisesta kerrotaan enemmän KTL:n valmisteilla olevassa menetelmäraportissa.

Verenpainemittareita oli alkuperäisen suunnitelman mukaan tarkoitus jättää 3 500 satunnaisesti valitulle 44–74-vuotiaalle haastatellulle. Suunnitelmasta poiketen mittareita jätettiin loppujen lopuksi 2 069 henkilölle. Tärkein yksittäinen syy (noin 80 %) siihen, miksi mittaria ei jätetty haastateltavalle, oli mittareiden saatavuusongelmat. Muita syitä olivat mittaamisesta kieltäytyminen sekä se, että haastateltu ei ollut tulossa terveystarkastukseen tai haastattelu oli tehty niin lähellä terveystarkastusta, että niiden väliin ei jäänyt riittävästi mittausaikaa. Jos haastattelun ja terveystarkastuksen väliin jäi vähemmän kuin neljä päivää eli yli puolet mittauksista olisi jäänyt tekemättä, mittaria ei jätetty haasta-

tellulle. Muutamat henkilöt eivät pystyneet mittaamaan verenpainettaan sairauden, kieliongelmien tai muiden vastaavien syiden takia.

Haastateltavien tavoittelu ja osallistumisasteen lisääminen

Haastattelijoille painotettiin tutkimuksen tärkeyttä ja heitä kehotettiin vielä tavallista aktiivisemmin tavoittelemaan tutkittavia haastattelun sopimiseksi. Haastateltavia yritettiin tavoitella eri aikoihin vuorokaudesta. Jos henkilöä ei useista yrityksistä huolimatta tavoitettu puhelimitse, haastattelija kävi hänen kotiosoitteessaan. Haastattelijat tekivät näitä käyntejä tarvittaessa vähintään kolme. Puhelimitse tapahtuvan tavoittelun määrää ei rajoitettu. Jos henkilö oli muuttanut otoksen poimimisen ja haastattelun välissä, haastattelija selvitti uuden osoitteen ja pyrki joko itse haastattelemaan tämän tai siirsi haastattelun lähempänä tutkittavaa asuvalle haastattelijalle.

Jos haastateltavaa ei tavoitettu lainkaan viimeistään 10 vuorokautta ennen terveystarkastusten loppumista kyseisellä paikkakunnalla, haastattelija merkitsi haastattelulomakkeelle tiedon, että tätä ei oltu tavoitettu. Sen jälkeen tieto siirrettiin haastattelijoiden tietojärjestelmän välityksellä KTL:n ajanvarausta hoitavaan yksikköön, josta tutkimushenkilölle lähetettiin vielä kutsu terveystarkastukseen. Jos aikaisemmin tavoittamaton henkilö ehti tulla terveystarkastukseen ennen niiden päättymistä, hänet pyrittiin myös haastattelemaan jälkikäteen. Jos henkilöä ei edelleenkään tavoitettu, Tilastokeskuksen haastattelija jatkoi tavoittelua kenttätyöajan (helmikuun 2001) loppuun asti.

Haastattelijoille painotettiin sitä, miten tärkeää oli, että kaikki tutkittavat saadaan haastateltua. Monet terveysongelmat, esimerkiksi vakavat toiminnanrajoitukset sekä mielenterveysongelmat, ovat selvästi yleisempiä niiden henkilöiden keskuudessa, jotka eivät halua osallistua tutkimukseen (ks. esim. Kattainen et al. 2003). Siksi tutkittavia yritettiin motivoida monin tavoin ja lisäksi haastattelijoiden katotyöskentelyä tehostettiin.

Kenttätyön aikana haastateltavia kohdehenkilöitä saatettiin vaihtaa haastattelijalta toiselle. Jo tavoitettu henkilö saatettiin siirtää toiselle haastattelijalle, jos hän ei alunperin ollut suostumassa tutkimukseen, mutta kieltäytyminen ei ollut aivan ehdoton. Esimerkiksi mieshaastattelijan sijasta haastateltavaa saattoi lähestyä seuraavalla kerralla naishaastattelija. Haastattelijan vaihto antaa haastateltavalle uuden mahdollisuuden osallistua tutkimukseen, mikäli häntä on edellisellä kerralla lähestytty kiireisenä tai muuten epäsopivana hetkenä, jolloin haastateltavan ja haastattelijan välille ei ole syntynyt positiivista pohjaa myöhemmälle yhteistyölle.

Haastattelusta kieltäytyneille henkilöille lähetettiin ns. kieltäytyneiden kirje, jossa motivointiin oli kiinnitetty erityistä huomiota koettamalla hyödyntää tutkimuksesta kieltäytyneiden mainitsemia syitä. Kirjeessä painotettiin tutkimuksen ainutlaatuisuutta sekä sen tärkeyttä terveydenhuollon suunnittelun taustatietona. Tämän jälkeen kieltäytyneeseen henkilöön otettiin uudelleen yhteyttä, ja yleensä myös vaihdettiin haastattelijaa. Lisäksi katotapauksia käsiteltiin haastattelijoiden alueryhmissä, joihin osallistui lähialueilla työskenteleviä haastattelijoita ja usein myös Tilastokeskuksen työnohjaaja.

Osa haastateltavista ei pystynyt osallistumaan Tilastokeskuksen haastatteluun ennen terveystarkastusta tai ei halunnut osallistua siihen ollenkaan. Tällöin KTL teki terveystarkastuksen yhteydessä lyhennetyn haastattelun. Jos haasta-

teltava ei halunnut osallistua haastatteluun eikä terveystarkastukseen, KTL pyrki saamaan puhelimitse edellisiä vaihtoehtoja lyhyemmän haastattelun, johon oli valittu joitakin tärkeimpiä kysymyksiä.

Suurin osa kohdehenkilöistä osallistui Tilastokeskuksen tekemään terveyshaastatteluun. Jos tutkittava ei pystynyt osallistumaan tutkimukseen ollenkaan sinä aikana, kun terveystarkastuksia pidettiin samalla paikkakunnalla, hänellä oli mahdollisuus käydä toisen paikkakunnan terveystarkastuksissa. Jos tämäkään ei ollut mahdollista esimerkiksi sairauden takia, terveystarkastus tehtiin kotona. Samalla näille henkilöille tehtiin lyhennetty haastattelu ja heille annettiin lyhennetyt versiot peruskyselystä ja infektiokyselystä (kyselyt 1 ja 2).

2.2 Sähköinen haastattelulomake ja tietoliikenne

Tilastokeskuksen haastattelijoiden tekemä terveyshaastattelu toteutettiin tietokoneavusteisena, kuten käytännöllisesti katsoen kaikki muutkin Tilastokeskuksen haastattelut. Tietokoneavusteiselle haastattelulle on ominaista, että haastattelija lukee kysymykset tietokoneen näytöltä ja tallettaa tiedot suoraan koneen muistiin. Paperilomaketta ei siis käytetä ollenkaan. Käyntihaastatteluissa haastattelijoilla on käytössään kannettavat tietokoneet.

Tilastokeskuksessa käytetään Alankomaiden tilastoviraston kehittämää Blaise-haastattelujärjestelmää, jonka oheen Tilastokeskuksessa on kehitetty kenttähaastatteluja tukeva tietojärjestelmä. Haastattelijoiden keräämät haastattelutiedot toimitetaan Tilastokeskukseen määrävälein, yleensä päivittäin, modeemien ja tähän tarkoitukseen suunnitellun tietoliikenneohjelmiston avulla. Tilastokeskuksessa haastattelijoiden lähettämät erilliset tiedostot kootaan yhdeksi tutkimustiedostoksi. Samalla kenttähaastatteluista tuotetaan päivittäinen seurantaraportti.

Kansanterveyslaitos (KTL) otti myös Blaise-ohjelmiston käyttöönsä Terveys 2000 -tutkimuksen alkaessa. Näin ollen Tilastokeskuksen haastattelijoiden keräämät tiedot voitiin toimittaa sellaisenaan Kansanterveyslaitokseen jatkokäsittelyä varten. Ainoa toimenpide, joka Tilastokeskuksessa tehtiin, oli tiedostomuunnos, joka johtui eri versioista. Tilastokeskuksessa oli Terveys 2000 -tutkimusta tehtäessä vielä käytössään Blaise-ohjelmiston DOS-versio, kun taas KTL otti käyttöönsä viimeisimmän Windows-version. Myös tutkimuksen kliinisen vaiheen tietojen keruu tehtiin suurelta osin Blaise-ohjelmiston avulla.

Sähköinen haastattelulomake poikkeaa huomattavasti paperilomakkeesta muutenkin kuin ulkoisten tekijöiden osalta. Sähköinen lomake on eräänlainen tietokoneohjelma, ja tätä kautta tiedonkeruussa on käytettävissä monia sellaisia piirteitä, joita ei ole paperilomakkeita käytettäessä. Kysymyksiä voidaan muotoilla dynaamisesti haastattelun edetessä lisäämällä niihin tekstiä aikaisemmista vastauksista tai muuten tiedossa olevista tiedoista. Näin kysymykset saadaan kohdennettua paremmin kuhunkin tilanteeseen sopivaksi. Lisäksi kysymysten kieliasu saadaan sellaiseksi, että se soveltuu käytettäväksi standardoidussa haastattelussa.

Ohjelmassa määritellään, missä järjestyksessä kysymykset esitetään. Kysymysjärjestys voi olla tarkasti asetettu tai se voi olla ehdollinen. Haarautuminen voi tapahtua aikaisempien vastausten tai vastaajan taustatietojen perusteella. Annettuja vastauksia voidaan tarkistaa monella tavalla haastattelun aikana. Tarkistukset voidaan kohdistaa vastausten arvoalueeseen. Hyväksyttävä arvoalue voidaan määritellä ehdolliseksi niin, että siinä otetaan huomioon oheistietoja. Tietojen tarkistaminen voidaan kohdistaa myös mahdolliseen ristiriitaan aikaisempien vastausten tai taustatietojen suhteen.

Sähköisen lomakkeen etuna on, että kysymysten muotoilun, ohjelmoidun kysymysjärjestyksen ja tarkistusten ansiosta tietokoneavusteisesti kerätty tieto on useimmiten oleellisesti virheettömämpää kuin aikaisemmin paperilomakkeilta tallennettu tieto. Käytännössä tämä tarkoittaa sitä, että kerättyä tietoa ei enää tarvitse mainittavasti tarkistaa, koska sellaiset toimenpiteet, joita paperilomakkeilta tallennetun tiedon tarkistamiseksi on tehty, tehdään sähköisessä lomakkeessa jo tiedon keruun yhteydessä.

Sähköisen tiedonkeruun tuloksena haastattelutiedot ovat nopeammin käytettävissä kuin paperilomakkeita käytettäessä, koska tiedot ovat valmiiksi tarkistettuja eikä erillistä tietojen tallennusta tarvita. Tätä tiedonkeruuprosessin nopeutta käytettiin hyväksi Terveys 2000 -tutkimuksessa ratkaisevalla tavalla.

Elektronisiin lomakkeisiin voidaan lisäksi liittää sellaisia ominaisuuksia, jotka eivät ole mahdollisia paperilomakkeita käytettäessä. Terveyshaastatteluissa sovellettiin muun muassa tietokoneavusteista koodausta monen kysymyksen yhteydessä. Haastattelijat koodasivat haastattelun yhteydessä vastaajan ammatin, syntymäpaikan, kotikunnan, sairaudet, ammattitaudit, vastaajan käyttämät lääkkeet ja hänelle tehdyt leikkaukset. Näin ollen jälkikäteen tehtävää erillistä koodausta ei tarvittu.

Sähköinen haastattelulomake poikkeaa perinteisestä paperilomakkeesta niin paljon, että sitä ei ole mahdollista kuvata tarkasti paperilomakkeena. Niinpä elektronisesta versiosta tehty tuloste antaa vain yleiskuvauksen haastattelusta (ks. www.ktl.fi/terveys2000). Täysin oikean käsityksen haastattelun kulusta ja kysymysten reitityksestä saa vain sähköisen lomakkeen kautta.

Tietojärjestelmä

Tilastokeskuksen kenttähaastatteluja tukeva tietojärjestelmä huolehtii haastateltavien yhteystietojen ja sähköisten lomakkeiden jakamisesta haastattelijoille sekä kerättyjen tietojen vastaanottamisesta ja yhdistämisestä. Tietojen siirto tehdään tietojärjestelmän osana toimivalla tietoliikenneohjelmistolla, jossa yhteys muodostetaan modeeminen välityksellä. Järjestelmän toimintaperiaate on karkeasti kuvattu kahdessa artikkelissa (Kuusela ja Parviainen, 1997; Kuusela 2001).

Terveys 2000 -tutkimuksessa haastattelijat lähettivät päivittäin keräämänsä haastattelutiedot Tilastokeskukseen, jossa eri haastattelijoiden tiedot yhdistettiin. Samalla tiedot muunnettiin KTL:n tarvitsemaan muotoon. Tilastokeskuksesta haastattelutiedot toimitettiin päivittäin tai joka toinen päivä KTL:een salakirjoitettuna sähköpostin liitetiedostona aikaleimalla varustettuna ja KTL kuittasi saamansa postin.

KTL:lla oli siis haastattelutiedot käytössään muutama päivä varsinaisen haastattelun jälkeen. KTL toimitti viidelle kiertävälle kliiniselle tutkimusyksi-

kölle haastattelijoiden sopimat tutkimusajat samoin kuin osan haastattelussa kerätyistä tiedoista. Tällä järjestelyllä kliininen tutkimus voitiin tehdä muutama viikko kenttähaastattelun jälkeen ja kenttähaastattelussa saatuja tietoja voitiin käyttää hyväksi kliinisessä tutkimuksessa. Terveys 2000 -tutkimuksen tietojärjestelmä on karkealla tasolla kuvattu erillisessä artikkelissa (Kuusela, Tanskanen, Virtala, 2000).

Kuvattu tietojärjestelmä, joka koostui Tilastokeskuksen kenttähaastattelujärjestelmästä, KTL:n tähän tutkimukseen suunnitellusta tietojärjestelmästä ja liikkuvien klinikoiden tietojärjestelmästä, vaikutti keskeisesti tutkimuksen tiedonkeruun onnistumiseen. Luultavasti tällainen toiminnan järjestely, tällä aikataululla, ei käytännössä olisi ollut mahdollista ilman edellä kuvattua tietojärjestelmää.

Ajanvaraustietojen välittäminen kenttäyksiköille

Terveystarkastuksia oli samanaikaisesti tekemässä viisi liikkuvaa tiimiä eri puolilla maata. Tiimit vaihtoivat paikkakuntaa etukäteen laaditun aikataulun mukaan. Haastattelut ja terveystarkastukset oli porrastettu siten, että niiden väliin ei jäänyt kovin pitkää aikaa ja että tutkittaville ei tullut kohtuuttoman pitkää matkaa terveystarkastukseen.

Kliininen tutkimus pyrittiin järjestämään nopeasti haastattelun jälkeen (2–3 viikkoa). Jotta liikkuvien klinikoiden toiminta saatiin mahdollisimman tehokkaaksi, ajanvaraus annettiin haastattelijoiden tehtäväksi haastattelun yhteydessä ja ajanvaraustiedot liitettiin muuhun tietoliikenteeseen.

KTL sai varatut ajat haastattelutietojen mukana parin päivän viiveellä haastattelun jälkeen. KTL jatkoi ajanvaraustiedot sähköisesti oikeille kenttäryhmille, jotka sijaitsivat eri puolilla maata. Näin jokaisella kenttäryhmällä oli tiedossaan merkityt aikavaraukset muutama päivää haastattelun jälkeen myös niiden paikkakuntien osalta, joilla ei kenttäryhmä vielä ollut käynyt. Tämä helpotti olennaisesti kenttäryhmien töiden suunnittelua.

Näin tehtynä ajanvarausjärjestelmästä tuli hyvin joustava, eikä se vaatinut mainittavasti lisäresursseja. Terveys 2000 -tutkimus oli hyvin nopeatempoinen eri paikkakunnilla liikkuvien klinikkojen vuoksi. Tutkimuksen eteneminen niin tiedonkeruun kuin ajanvarauksenkin osalta oli keskeisellä tavalla riippuvainen tietojärjestelmän toiminnasta. Jälkikäteen arvioiden tulos oli onnistunut.

2.3 Tiedonkeruuprosessin arviointi

Terveys 2000 -tutkimuksen tiedonkeruun suunnittelu oli vaativa tehtävä, koska tiedonkeruu koostui lukuisista erilaisista haastatteluista ja mittauksista, joita toteuttivat useat erilliset yksiköt ja erilaisen taustan omaavat ihmiset. Tiedonkeruun suunnitteluun osallistui suuri joukko asiantuntijoita muun muassa Kansanterveyslaitoksesta ja Tilastokeskuksesta. Tilastokeskuksen vastuulla oli terveyshaastattelujen käytännön toteutus.

Liikkuvien klinikoiden useasta eri osasta koostuvan ja eri puolilla maata toteutetun sähköisen tiedonkeruun järjestäminen vaati tavanomaista runsaammin suunnittelua ja yhteistyötä, koska sähköinen keruutapa oli tuttu ja rutiini-

käytössä vain Tilastokeskuksessa. Tutkimusta ei kuitenkaan olisi voitu toteuttaa tässä laajuudessa ja tällä aikataululla ilman nyt käytettyä, suhteellisen monimutkaista teknistä järjestelmää sekä suurta joukkoa suunnittelijoita, tutkimushenkilökuntaa ja haastattelijoita. Kahden pilottitutkimuksen kautta ennen varsinaisen tiedonkeruun käynnistymistä saatiin hyödyllistä palautetta, jonka avulla tietojärjestelmän ja tiedonkeruun toimivuutta voitiin parantaa.

Yleisesti ottaen tiedonkeruun arvioitiin onnistuneen hyvin, ja se toteutui suunnitellussa aikataulussa. Sähköiselle tiedonkeruulle on ominaista, että tiedonkeruuta edeltävä suunnittelu vaatii paljon enemmän aikaa kuin paperilomakkeisiin perustuva tiedonkeruu. Tästä johtuen muutamat tiedonkeruuta edeltäneet viiveet aiheuttivat kiirettä. Erityisesti haastattelulomakkeen ohjelmoinnissa jouduttiin toimimaan suunniteltua tiukemmassa aikataulussa sisällön hitaan valmistumisen vuoksi. Syynä oli ilmeisesti se, että kaikki suunnitteluun osallistuvat eivät olleet tottuneet sähköisen tiedonkeruun edellyttämään tutkimusaikatauluun. Niinpä sisältömuutoksia jouduttiin tekemään ja virheitä korjaamaan aivan ohjelmoinnin ja testauksen loppuun asti.

Olosuhteisiin nähden ohjelmointi onnistui erittäin hyvin. Joitakin vähäisiä virheitä kuitenkin havaittiin vasta haastatteluiden käynnistyttyä tai niiden jälkeen. Näin käy usein monimutkaisten sähköisten lomakkeiden kohdalla, koska niiden täydellinen testaaminen muun muassa monien sisäänrakennettujen polkurakenteiden vuoksi on erittäin vaativa tehtävä. Virheet eivät kuitenkaan vaikuttaneet haastattelulomakkeen tekniseen toimintaan, ja siksi haastattelut pystyttiin toteuttamaan suunnitellusti.

Yksi esimerkki virheistä on kysymys "Onko sairaudesta haittaa työssä?", joka kysyttiin vain työssä olevilta henkilöiltä. Alkuperäinen tarkoitus oli esittää kysymys myös omaa kotitaloutta hoitavilta sekä sairaus-, äitiys-, isyys- tai vanhempainlomalla oleville, jotka olivat olleet työssä haastattelua edeltäneiden 12 kuukauden aikana ja joilla on jokin sairaus.

Haastattelussa käytettiin useita erilaisia koodistoja. Vakiintuneemmassa käytössä olevat koodistot toimivat parhaiten, muokattu kuntakoodisto mukaan luettuna. Sen sijaan varsinkin leikkauskoodisto toimi haastattelijoiden mielestä huonosti käytännön haastattelutyössä. Koodistossa oli sisällöllisiä ja rakenteellisia puutteita, ja paikoin sen kieli ei vastannut haastateltavien käyttämää kieltä (esimerkiksi umpilisäke vs. umpisuoli). Terveystarkastuksiin liittyvissä ruotsinkielisissä kysymyksissä oli lisäksi vastaajille tuntemattomia sanoja (palpation av brösten, prostatapalpation, toxemi).

Tilastokeskuksen haastattelijat koulutettiin elokuussa 2000. Osalla haastattelijoista työ alkoi heti koulutuksen jälkeen, mutta osa joutui terveystarkastusaikataulujen takia odottamaan työn alkamista seuraavan vuoden puolelle. Tämä saattoi aiheuttaa jonkin verran ohjeiden muistamattomuutta, varsinkin kun osa ohjeista ja korjauksista tuli useina erillisinä lähetyksinä useiden kuukausien aikana. Toisaalta myöhemmin työnsä aloittaneet haastattelijat saattoivat hyötyä siitä, että käytännöt olivat ehtineet vakiintua ja ohjeita oli täydennetty kenttätyössä tehtyjen havaintojen perusteella.

Haastattelu sisälsi monia kysymyksiä, joihin vastaaminen vaati muistamista joskus pitkälle taaksepäin. Monilla vastaajilla olikin muistamisongelmia esimerkiksi työhistoriassa, eräissä sairauskysymyksissä ja vanhemmilla vastaajilla myös joissakin hedelmällisyyteen ja lisääntymiseen liittyvissä kysymyksissä.

Haastattelijoiden työtä vaikeuttivat lisäksi verenpainemittareiden toimitusongelmat sekä kenttätyön aikana toimitetut lukuisat lisäohjeet, jotka vaativat jatkuvaa tietojen päivittämistä. Tällaisia hankaluuksia tulee väistämättä suuressa ja monimutkaisessa tutkimushankkeessa, jossa tietoa tuotetaan, päivitetään ja siirretään useissa eri organisaatiossa työskentelevien ihmisten kesken.

Tutkimuksesta tiedotettiin valtakunnallisissa uutislähetyksissä ja myöhemmin terveystarkastusten alkaessa paikallislehdissä. Tutkimushankkeen saama julkisuus tiedotusvälineissä lisäsi haastateltavien myönteistä suhtautumista tutkimukseen. Tosin haastattelijoiden työn kannalta paikallinen tiedottaminen oli toisinaan myöhässä, koska se käynnistyi vasta terveystarkastusten alkaessa, jolloin haastatteluita oli jo tehty. Riittävän ajoissa alkanut tiedottaminen motivoi haastattelijoiden arvion mukaan haastatteluun ja samalla koko tutkimukseen, osallistumista. (T. Nieminen, 2003).

Haastatteluita tehtiin seitsemän kuukauden aikana. Taulukossa 2.1. on tehtyjen haastatteluiden määrät kuukausittain eri yliopistosairaala- eli miljoonapiireissä.

Taulukko 2.1.
Kotihaastattelut kuukausittain eri miljoonapiireissä.

	MILJOONAPIIRIT					
Kuukausi/Vuosi	**HYKS**	**TYKS**	**TAYS**	**KYS**	**OYS**	**Yhteensä**
08 / 2000	7	25	40	11	11	94
%	0,3	2,5	2,5	0,9	1,1	1,4
09 / 2000	591	348	639	319	410	2 307
%	27,0	35,3	40,4	25,9	41,0	33,0
10 / 2000	659	268	416	271	213	1 827
%	30,1	27,2	26,3	22,0	21,3	26,2
11 / 2000	398	190	183	285	98	1 154
%	18,2	19,3	11,6	23,1	9,8	16,5
12 / 2000	207	68	95	113	76	559
%	9,5	6,9	6,0	9,2	7,6	8,0
01 / 2001	284	76	179	192	167	898
%	13,0	7,7	11,3	15,6	16,7	12,9
02 / 2001	44	10	28	41	24	147
%	2,0	1,0	1,8	3,3	2,4	2,1
Yhteensä	**2 190**	**985**	**1 580**	**1 232**	**999**	**6 986**
%	**100**	**100**	**100**	**100**	**100**	**100**

Tiedonkeruu alkoi elokuun 2000 loppupuolella, joten elokuussa haastatteluita ei tehty vielä kovin monta. Myös kenttätyön päättyessä helmikuun 2001 lopussa haastatteluiden määrä oli pienempi kuin edeltävinä kuukausina. Tutkimus aloitettiin isoista yliopistosairaalakaupungeista, joissa otoskoko oli suurin. Siitä syystä haastatteluja on tehty eniten syys–lokakuussa 2000. Kussakin miljoonapiirissä tutkituista henkilöistä suurin osa haastateltiin syksyllä 2000. Tehtyjen

haastatteluiden suhteellinen osuus eri kuukausina kuitenkin vaihtelee huomattavasti miljoonapiireittäin.

Terveys 2000 -tutkimuksessa tavoiteltiin keskeisten tulosmuuttujien osalta vertailukelpoisuutta Mini-Suomi-tutkimuksen kanssa. Tulosten vertailussa on kuitenkin huomioitava tiedonkeruun organisoinnin erot tutkimusten välillä: Mini-Suomi-tutkimus toteutettiin terveydenhoitajien tekemien paperilomakehaastatteluiden (PAPI) avulla, kun taas Terveys 2000 -tutkimuksessa haastatteluja tekivät Tilastokeskuksen ammattihaastattelijat tietokoneavusteisina käyntihaastatteluina (CAPI). Erot tiedonkeruussa ja haastattelijoiden työorientaatiossa saattavat joiltakin osin vaikuttaa tutkimusten vertailukelpoisuuteen. Lehtonen (1996) ja Nieminen (1997) ovat verranneet ammattihaastattelijoiden ja terveydenhoitajien tekemiä haastatteluja ja raportoineet havaitsemistaan poikkeamista terveystutkimusten eräiden osa-alueiden tuloksissa.

3 Otanta-asetelma

Johanna Laiho ja Kari Djerf

Terveys 2000 -tutkimuksen otanta-asetelma perustuu kaksiasteiseen ositettuun otantaan, jonka otantakehikkona käytettiin Kansaneläkelaitoksessa (KELA) ylläpidettävää sosiaalivakuutettujen henkilörekisteriä. Tässä luvussa arvioidaan otantakehikon sopivuutta ja peittävyyttä sekä kuvataan yksityiskohtaisesti tutkimuksen otanta-asetelma ja lopullisen otoksen muodostuminen.

3.1 Otantakehikko

Tavoiteperusjoukon määritys

Terveys 2000 -tutkimuksen tavoiteperusjoukoksi määriteltiin Manner-Suomessa vakinaisesti asuva 18 vuotta täyttänyt aikuisväestö. Tavoiteperusjoukkoon kuului kotitalousväestön lisäksi myös laitosväestö. Alueellisen rajauksen vuoksi koko Ahvenanmaan maakunta ja ulkosaariston ilman suoraa tieyhteyttä olevat kunnat eli Hailuoto, Houtskari, Iniö, Korppoo, Nauvo ja Velkua suljettiin otoskehikon ulkopuolelle.

Otantakehikko ja sen laatu

Väestötietojärjestelmää (VTJ) ja siihen perustuvia erillisiä väestötiedostoja käytetään pääasiallisena otantakehikkona kaikissa Suomen Virallisen Tilaston henkilö- ja kotitalousotoksissa. VTJ kattaa kaikki Suomessa vakituisesti asuvat henkilöt. Tietojärjestelmä sisältää alue-, paikka- ja osoitetietojen lisäksi demografista tietoa henkilöistä ja asuntokunnista, kuten esimerkiksi henkilön iän, sukupuolen, äidinkielen sekä asuntokunnan koon. Tämän lisäksi järjestelemä sisältää henkilökohtaisen identifiointitunnuksen eli henkilötunnuksen, mikä mahdollistaa taustatiedon yhdistämisen tilastotoimen tarkoituksiin muista rekistereistä ja hallinnollisista tietolähteistä.

Terveys 2000 -tutkimuksen otos poimittiin Kansaneläkelaitoksessa, ja otantakehikkona käytettiin KELAn sosiaalivakuutettujen henkilörekisteritietokantaa, jota päivitetään jatkuvasti sekä väestötietojärjestelmästä että KELAn paikallistoimistoissa. Tietokantaan pystyttiin myös yhdistämään terveystutkimukselle olennainen terveyskeskuspiiri- ja miljoonapiiriluokitus. Sosiaalivakuutettujen henkilörekisteritietokannan kattavuus määrittää tutkimuksen kohdeperusjoukon.

Kehikon peittävyys, täydellisyys, ajantasaisuus, tietosisältö ja tietojen tarkkuus ovat kriittisiä tekijöitä kehikon sopivuudelle tilastotutkimuksen käyttöön. Kehikon tulee myös sisältää valitun otantamenetelmän tarvitsemaa, luotettavaa ja mahdollisimman ajantasaista lisäinformaatiota, esimerkiksi otoksen osituksseen. Otantakehikoiden arviointia ja sopivien otantamenetelmien valintaa Suo-

men tilastotoimen toimintakehyksessä on kuvattu tarkemmin muun muassa Laihon ja Hietaniemen (2002) toimittamassa Laatua tilastoissa -käsikirjassa.

Koska KELAn sosiaalivakuutettujen henkilörekisteritietokanta ja väestötietojärjestelmä ovat kattavuudeltaan ja ajantasaisuudeltaan yhteneviä, on tässä yhteydessä käytetty kehikon laadun arviointiin saatavilla olevia VTJ:n luotettavuustutkimuksia ja -arvioita. Tilastokeskuksessa (TK) arvioidaan säännöllisesti VTJ:n luotettavuutta: viimeisimmät luotettavuustutkimukset on tehty Ylitalon (2002) ja M. Niemisen (2003) toimesta. Näissä luotettavuustutkimuksissa on estimoitu virheellisten osoitteiden osuudeksi 1,1 prosenttia vuonna 2002 ja 0,8 prosenttia vuonna 2003. Molemmissa luotettavuustutkimuksissa osoitetieto puuttui 0,2 prosentilta tutkimukseen poimituista henkilöistä.

VTJ:n päivittäminen on yleisesti ottaen nopeaa ja kattavaa, minkä ansiosta sen tiedot ovat myös hyvin ajantasaisia. Ruotsalainen (2002) arvioi, että noin 3 prosenttia väestötiedoista voi sisältää jotain virhettä. Tähän arvioon sisältyvät myös epätäydelliset osoitetiedot. Useissa tapauksissa TK:n kokeneet haastattelijat pystyvät selvittämään täydellisen osoitetiedon. Lisäksi haastattelijat ovat harjaantuneet jäljittämään ihmisiä, joiden osoitetiedot ovat vanhentuneet.

Otostiedoston rajaukset

Otoksen muodostamisessa hyödynnettiin KELAn sosiaalivakuutettujen henkilörekisteritietokannan sisältämiä tunnustietoja: henkilötunnuksesta johdettua henkilön ikää ja kotipaikkatunnuksesta johdettua kotikuntatunnusta, jonka avulla tiedostoon yhdistettiin terveyskeskuspiirien ja miljoonapiirien tunnukset sekä niiden luokitusavainta suhteessa kuntajakoon.

Henkilörekisteritietokannasta muodostettiin Kansaneläkelaitoksessa otostiedosto, johon poimittiin kaikki maassa asuvat 18 vuotta (1.7.2000) täyttäneet henkilöt. Otostiedostosta pyrittiin poistamaan mahdollisimman paljon ylipeittoa jo etukäteen, kuten esimerkiksi:

- tilapäisesti ulkomailla asuvat,
- henkilöt, joiden kotipaikkakunta on ulkomailla,
- ulkomailla asuvat diplomaatit ja lähetyshenkilökunta, (ei vakituista asuntoa Suomessa),
- henkilöt, joiden olinpaikasta ei ole tietoa.

Tietoturvasyistä otostiedostosta poistettiin myös niin kutsutut luovutuskieltotapaukset eli henkilöt, joiden olinpaikka on määritelty ehdottomasti salassapidettäväksi. Luovutuskieltotapaukset on merkitty erikseen väestötietojärjestelmän tiedostoihin.

Ylipeitto

Henkilörekisterin ylipeiton muodostavat kuolleet tai maasta muuttaneet henkilöt. Ylipeittoon kuuluvat myös ne henkilöt, joiden tapahtumia ei ole päivitetty rekisteritietoihin ennen otoksen poimintaa. Maasta muuttaneet kuuluvat uuden kohdemaansa väestöön, eikä heitä enää lueta mukaan Manner-Suomen väestöön. Terveys 2000 -tutkimuksessa ylipeittoon kuului 78 henkilöä alkuperäisestä otoksesta.

Yleensä suurin osa ylipeitosta pystytään havaitsemaan ennen kenttätyövaihetta ja sen aikana.

Otoksen poiminnan jälkeen ennen varsinaista kenttätyövaihetta henkilöiden osoitetiedot päivitetään mahdollisimman uusilla tiedoilla väestötietojärjestelmästä, jolloin havaitaan mahdolliset uudet ylipeittotapaukset. Tämän vuoksi tehdään käsitteellinen ero alkuperäisen eli brutto-otoksen ja lopullisen otoksen eli netto-otoksen välillä: netto-otos = brutto-otos - ylipeitto.

Alipeitto

Otantakehikon ylipeitto voidaan mitata, mutta alipeittoa on vaikeampi arvioida (Djerf, 2000). Otantakehikon alipeitto muodostuu niistä henkilöistä, jotka kuuluvat tavoiteperusjoukkoon, mutta joiden tietoja ei ole eri syistä otoksen poimintakehikossa otoksen poimintahetkenä. Käytännössä alipeittoa aiheutuu maahan muuttaneista henkilöistä, joiden muuttotapahtumaa ei ole päivitetty ajoissa KELAn sosiaalivakuutettujen henkilörekisteriin. Terveys 2000 -tutkimuksessa on alueellisen ryvästymisen vuoksi huomioitava myös maassamuuton vaikutus peittovirheeseen.

Maahanmuuton suhteen alipeittovirhe syntyi ainoastaan niistä muuttajista, jotka ovat muuttaneet tutkimukseen valittuihin terveyskeskuspiireihin ennen otoksen poimintaa ja joiden tietoja ei ole päivitetty. Maassamuuton suhteen alipeittoa syntyi niistä muuttajista, jotka muuttivat tutkimukseen kuulumattomasta terveyskeskuspiiristä tutkimukseen valittuun terveyskeskuspiiriin ja tekivät muuttoilmoituksen liian myöhään tai jättivät sen kokonaan tekemättä. Tämän lisäksi on huomioitava, että rekisteröitymättömän maahanmuuton aiheuttama alipeitto voi vaikuttaa hieman koko Manner-Suomen tasolla väestöestimaatteihin, joihin otantatutkimuksen tulokset pyritään yleistämään.

Alipeiton suuruutta on hyvin vaikeata estimoida, koska kyse on vaihtelevasti muutaman päivän tai viikkojen viipeestä muuttotapahtuman päivityksestä väestötietojärjestelmään, josta tiedot päivitetään sosiaalivakuutettujen henkilörekisteriin. Alipeiton arvioidaan olevan hyvin vähäistä, sillä muuttoliike, joka kohdistuu maan ulkopuolelta tai maan sisältä valittuihin terveyskeskuspiireihin, ei ole ollut poikkeuksellisen suurta vuonna 2000.

Alipeiton aiheuttama harha voi olla kokonaisuudessaan hyvin pieni, mutta toisaalta se voi olla painottunut nuoriin aikuisiin, jotka muuttavat lukukausien välillä opiskelu- ja työmahdollisuuksien perässä. Tässä yhteydessä on huomattava, että Terveys 2000 -tutkimuksen otoksen poiminta ajoittui tilanteeseen 31.7.2000. Alipeiton vaikutusta on kuitenkin hyvin vaikea arvioida, etenkin kun muuttoliikkeen tekijät voivat vaihdella alueittain hyvin paljon. Terveys 2000 -tutkimuksen otos käsitti muun muassa Suomen 15 suurinta kaupunkia, joista osa oli muuttovoitto- ja osa muuttotappiokaupunkeja vuonna 2000.

3.2 Otanta-asetelman kuvaus

Otantamenetelmä

Terveys 2000 -tutkimukseen suunniteltiin Tilastokeskuksessa kaksiasteinen ositettua ryväsotantaa hyödyntävä otanta-asetelma. Otos poimittiin tämän otantasuunnitelman mukaisesti Kansaneläkelaitoksessa sosiaalivakuutettujen henkilörekisteritietokannasta. Tarkistusten jälkeen brutto-otoksessa oli 30 vuotta täyttäneitä yhteensä 8 028 henkilöä, sekä 18–29-vuotiaita nuoria aikuisia 1 894 henkilöä.

Moniasteisen otannan osituksen tavoitteena oli poimia maantieteellisesti koko Manner-Suomen väestöstä edustava otos. Ryvästyksen tarkoituksena oli puolestaan parantaa otoksen kustannustehokkuutta kenttätyön näkökulmasta, erityisesti terveystarkastusten eli kliinisten tiedonkeruupisteiden organisoinnin helpottamiseksi. Ryvästyksen olennaisia kriteereitä olivat hyvät maakulkuyhteydet kliiniseen tiedonkeruupisteeseen sekä se, että matkaetäisyydet eivät tule liian pitkiksi tutkittaville henkilöille. Tämän lisäksi otanta-asetelmaa suunniteltaessa huomioitiin se, että haastattelijakohtaiset ja erityisesti kliinisen kenttätyövaiheen kenttäryhmäkohtaiset työmäärät tuli olla allokoitavissa optimaalisesti. Rypäiden poiminnassa noudatettiin normaaleja todennäköisyysotannan menetelmiä.

Osittaminen ja rypäät

Otoskehikko ositettiin alueellisesti terveydenhuollon viiden miljoonapiirin eli Helsingin yliopistollisen keskussairaalapiirin (HYKS), Turun yliopistollisen keskussairaalapiirin (TYKS), Tampereen yliopistollisen keskussairaalapiirin (TAYS), Kuopion yliopistollisen keskussairaalapiirin (KYS) ja Oulun yliopistollisen keskussairaalapiirin (OYS) mukaan käyttäen suhteellista kiintiöintiä väestömäärään suhteutettuna. Alueellisten ositteiden sisällä rypäänä on terveyskeskuspiiri, joita on kaikkiaan 249.

Otannan ensimmäisessä asteessa poimittiin 80 terveyskeskuspiiriä 249:stä, ja toisessa asteessa poimittiin kohdehenkilöt valituista terveyskeskuspiireistä. Jokainen osite jaettiin edelleen kahtia kussakin miljoonapiirissä seuraavasti. Koko maan 15 asukasluvultaan suurimman kaupungin terveyskeskuspiirit poimittiin otokseen todennäköisyydellä 1. Niissä otoskoko on suoraan suhteessa väestön määrään. Loppuotos eli yhteensä 65 terveyskeskuspiiriä poimittiin kussakin ositteessa systemaattisella PPS-otannalla eli suhteellisella kiintiöintimenettelyllä käyttäen kokomuuttujana asukaslukua.

Alueluokitusmuutokset

Kuntajako, terveyskeskuspiiri- ja miljoonapiiriluokitukset eivät ole täysin yhtenäisiä. Jotta terveyskeskuspiirien rajoja ei rikottaisi, tehtiin ennen poimintaa seuraavat alueluokitusmuutokset miljoonapiiritasolla:

- HYKS-miljoonapiiriin kuuluva Lavian kunta muodostaa yhdessä TAYS-miljoonapiiriin kuuluvien Suodenniemen ja Kiikoisten kuntien kanssa Lavian terveyskeskuspiirin. Tässä tutkimuksessa Lavian kunnan katsotaan kuuluvan TAYS-miljoonapiiriin.

- HYKS-miljoonapiiriin kuuluvat kunnat Myrskylä ja Pukkila muodostavat yhdessä TAYS-miljoonapiiriin kuuluvien Artjärven ja Orimattilan kuntien kanssa Orimattilan terveyskeskuspiirin. Tässä tutkimuksessa Myrskylän ja Pukkilan katsotaan kuuluvan TAYS-miljoonapiiriin.

Henkilöotoksen kiintiöinti ositteisiin

Otannan toisessa vaiheessa varsinainen henkilöpoiminta tehtiin ensimmäisessä vaiheessa poimituista 80 terveyskeskuspiiristä. Henkilöpoimintaan käytettiin systemaattista otantaa, jossa kunkin terveyskeskuspiirin väestö lajiteltiin iän mukaan. Todennäköisyydellä 1 poimittujen suurten kaupunkien tapauksessa otoskoko suhteutettiin väestön määrään, joten otanta oli yksiasteinen. Sen sijaan muissa rypäissä otoskoot laskettiin taulukossa 3.1. esitetyn asetelman avulla siten, että ositteiden otoskoko vastasi suhteellisen kiintiöinnin vaatimusta. Näiden rypäiden otanta oli kaksiasteinen.

Otos jaettiin tutkimusasetelman vuoksi kahteen ryhmään, joista 31.7.2000 mennessä 30 vuotta täyttäneille kohdistettiin koko Terveys 2000 -tutkimus terveyshaastattelu-, kyselylomake- ja terveystarkastusosioineen. Nuorille aikuisille (18–29 vuotta täyttäneille) suunnattiin vain haastattelu- ja kyselytutkimus. Aromaa ja Koskinen (2002) ovat esittäneet tarkan kuvauksen Terveys 2000 -tutkimuksen tutkimusasetelmasta.

Tutkimuksen pääryhmässä (30 vuotta täyttäneet) pienin ryväskohtainen otoskoko oli 50 ja suurin 100. Jokaisessa terveyskeskuspiirissä 80 vuotta täyttäneitä poimittiin kaksinkertaisella todennäköisyydellä, jotta vanhuksia saataisiin riittävästi Terveys 2000 -tutkimukseen. Nuorille aikuisille tarkoitetussa ryhmässä pienin otoskoko rypäässä oli 10 ja suurin 25.

Taulukko 3.1.
Poimittu ja odotettu otos kohdehenkilön iän ja miljoonapiirin mukaan

	Kohde-väestö	Rypäiden lkm	Odotettu otoskoko 18+	Poimittu otoskoko 18+	Odotetusta otoksesta 30+	Poimitusta otoksesta 30+	Aluerypään poiminta-todennäköisyys
1. HYKS	**1 305 804**	**16**	**3 278**	**3 436**	**2 620**	**2 811**	
1A. Rypäät 1–5:	822 181	5	2 064	2 161	1 650	1 716	
Helsinki	447 064	1	1 122	1 173	900	918	1
Espoo	153 597	1	386	405	309	316	1
Vantaa	131 252	1	330	346	260	283	1
Kotka	44 688	1	112	116	90	101	1
Lappeenranta	45 580	1	114	121	90	98	1
1B. Rypäät 6–16: tk-piirit	483 513	11	1 214	1 275	970	1 095	PPS
Odotettu otoskoko per ryväs			*110*		*88*		
2. TYKS	**533 310**	**16**	**1 339**	**1 412**	**1 070**	**1 178**	
2A. Rypäät 17–18:	201 053	2	505	534	400	428	
Turku	140 217	1	352	373	280	288	1
Pori	60 836	1	153	161	120	140	1
2B. Rypäät 19–32: tk-piirit	332 257	14	834	878	670	750	PPS
Odotettu otoskoko per ryväs			*60*		*50*		
3. TAYS	**925 769**	**16**	**2 324**	**2 441**	**1 860**	**2 046**	
3A. Rypäät 33–36:	326 945	4	821	861	656	719	
Tampere	154 467	1	388	406	310	334	1
Lahti	77 087	1	193	202	150	178	1
Vaasa	44 832	1	113	118	90	90	1
Hämeenlinnan tk-piirit	50 559	1	127	135	101	117	1
3B. Rypäät 37–48: tk-piirit	598 824	12	1 503	1 580	1 200	1 327	PPS
Odotettu otoskoko per ryväs			*125*		*100*		
4. KYS	**675 381**	**16**	**1 695**	**1 777**	**1 360**	**1 481**	
4A. Rypäät 49–51:	168 790	3	424	445	340	330	
Kuopio	67 354	1	169	175	140	131	1
Jyväskylä	61 160	1	154	164	120	115	1
Joensuu	40 276	1	101	106	80	84	1
4B. Rypäät 52–64: tk-piirit	506 591	13	1 271	1 332	1 020	1 151	PPS
Odotettu otoskoko per ryväs			*98*		*80*		
5. OYS	**543 676**	**16**	**165**	**1 426**	**1 090**	**1 153**	
5A. Ryväs 65, Oulu	89 537	1	225	240	180	181	1
5B. Rypäät 66–80: tk-piirit	454 139	15	1 140	1 186	910	972	PPS
Odotettu otoskoko per ryväs			*76*		*60*		

Poiminta miljoonapiirien sisällä kahdesta ositteesta:
A. Itse-edustavat ositteet (rypäät, joilla poimintatn=1)
B. Muut terveyskeskuspiirirypäät (poiminta PPS-otannalla, Probability Proportional to Size)

Rypäiden muodostamisen arviointi

Toteutunut otos oli lähellä odotettua: erot aiheutuvat 80-vuotiaiden kaksinkertaisesta sisältymistodennäköisyydestä ja otantakehikossa tapahtuneista muutoksista. Otos on osittain kohtuullisen hyvin itsepainottuva, mutta otosrypäiden luku on melko pieni. Taulukossa 3.2. on esitetty 30 vuotta täyttäneiden ikä- ja sukupuolijakauma netto-otokselle (painottamaton jakauma) sekä sen edustavuus perusjoukossa (painotettu jakauma).

Taulukko 3.2.
Poimittu otos ja sen edustavuus perusjoukossa kohdehenkilön iän ja sukupuolen mukaan

	Netto-otos			Edustavuus perusjoukossa		
Ikä	**Naisia**	**Miehiä**	**Kaikki**	**Naisia**	**Miehiä**	**Kaikki**
30–39	847	814	1 661	362 503	348 338	710 841
40–49	941	902	1 843	405 353	391 809	797 162
50–59	822	846	1 668	352 432	366 475	718 907
60–69	591	519	1 110	255 403	224 934	480 337
70–79	533	317	850	231 084	140 247	371 331
80+	615	203	818	131 792	44 311	176103
Yhteensä	**4 349**	**3 601**	**7 950**	**1 738 567**	**1 516 114**	**3 254 681**

4 Katoanalyysi

Johanna Laiho

4.1 Kato ja puuttuva tieto virhelähteenä

Kyselytutkimuksissa puuttuvaa tietoa eli vastauskatoa syntyy eri syistä. Yksikkökatoa aiheutuu muun muassa tutkimukseen valittujen henkilöiden tavoittamattomuudesta, tutkimuksesta kieltäytymisestä, sairastamisesta ja haastattelukielen taidon puutteista. Osa hyväksytysti vastanneista voi myös jättää vastaamatta tai osallistumatta tiettyihin tutkimuksen osiin ja/tai kysymyksiin. Tätä jälkimmäistä puuttuvan tiedon tyyppiä kutsutaan eräkadoksi.

Vastauskato voi aiheuttaa harhaa lopullisiin estimaatteihin (Cochran, 1963; Kish, 1965). Lisäksi vastanneiden ja vastausten määrän pieneneminen vähentää estimaattien tarkkuutta. Vastauskato voi näin ollen vähentää aineiston luotettavuutta ja tulosten yleistettävyyttä. Tutkimusaineistojen katoanalyysi tuottaa tärkeätä tietoa aineiston käytettävyydestä eri käyttötarkoituksiin. Yksikkö- ja eräkadon vaikutusten tunnistaminen ja korjaaminen on erittäin tärkeätä kyselytutkimusten teettäjille, tekijöille ja lopullisten tutkimusaineistojen käyttäjille. Kadon mahdollisia vaikutuksia otantatutkimusten lopullisiin estimaatteihin on tutkittu jo suhteellisen pitkään (Smith, 2002; Hansen ja Hurwitz, 1946; Politz ja Simmons, 1949). Kuitenkin Groves et al. (1992) painottavat vasta Demingin (1953) ja Hansenin et al. (1953) aloittaneen systemaattisen katoanalyysin ja kadon vaikutusta korvaavien menetelmien kehityksen.

Tässä luvussa tarkastellaan Terveys 2000 -tutkimuksen 30 vuotta täyttäneiden pitkän tai lyhyen terveyshaastatteluosion yksikkökatoa. Hyväksyttyjen vastausten määritys ja tarkistus on tehty Kansanterveyslaitoksen (KTL) toimesta. Luvussa esitetään pääosin painotettuja vastausosuuksia, jotka yleistävät kadon vaikutuksen koko väestön tason estimaatteihin. Kun vastausosuus on painotettu henkilöiden alkuperäisten sisältymistodennäköisyyksien käänteisluvulla, vastausosuus yleistää tutkimuksen kattavuuden koko kohdeperusjoukon tasolle ja kuvaa sitä osuutta populaatiosta, jota haastatellut kohdehenkilöt edustavat (Laiho, 2002a; Platek and Gray, 1986). Terveys 2000 -tutkimuksen painotus selitetään tarkemmin seuraavassa luvussa 5. Painottamaton vastausosuus kuvaisi pelkästään vastanneiden osuutta otoksen kohdeperusjoukkoon kuuluvista henkilöistä eli toisin sanoen kenttätyön onnistumista sille annettujen resurssien ja aikarajojen puitteissa. Myös tätä pyritään arvioimaan tässä luvussa. Kuvaileva katoanalyysi on tehty pääasiallisesti sukupuolen mukaan, koska miesten ja naisten vastauskäyttäytyminen voi poiketa toisistaan eri taustatekijöiden mukaan.

Terveys 2000 -tutkimuksen katoa tarkastellaan sen netto-otoksesta, joka saadaan poistamalla ylipeitto alkuperäisestä brutto-otoksesta. Ylipeitto kuvaa kehikkovirhettä, ja sitä syntyi Terveys 2000 -tutkimuksessa lähinnä kuolemien ja maastamuuton seurauksena. Otostiedot tarkistetaan aina ennen varsinaista kenttätyövaihetta. Tästä huolimatta ylipeittoa syntyy aina hieman johtuen otantakehikon väestötietojen päivittämisen viiveistä sekä toisaalta otoksen tarkista-

misen ja varsinaisen haastattelupäivän välille sijoittuvista muutoksista. Terveys 2000 -tutkimuksessa ylipeittoon kuului 78 henkilöä alkuperäisestä otoksesta.

Taulukossa 4.1. on esitetty terveyshaastatteluaineiston vastauskato kadon syyn mukaan netto-otokselle (painottamaton kato-osuus) sekä tarkasteltu kadon edustavuutta perusjoukossa (painotettu kato-osuus). Tarkasteltaessa haastattelututkimuksen onnistumista käytetään usein vastausosuutta kato-osuuden sijasta. Painottamaton vastausosuus kuvaa haastattelijoiden kenttätyön onnistumista, ja painotettu vastausosuus tulosten yleistettävyydestä tavoiteperusjoukkoon. Terveys 2000 -tutkimuksen 89,4 prosentin painotettu vastausosuus kertoo sen kohdeväestön osuuden, joka on edustettuna lopullisessa tutkimusaineistossa. Painottamaton 89,1 prosentin vastausosuus puolestaan mittaa hyväksytysti tutkimukseen haastateltujen henkilöiden osuutta alkuperäisestä netto-otoksesta. Tässä tutkimuksessa painotetut ja painottamattomat vastausosuudet poikkeavat vain vähän toisistaan.

Otantatutkimuksen aihe voi vaikuttaa saavutettuun vastausosuuteen. Myös pääsy terveystarkistukseen saattoi lisätä kohdehenkilöiden osallistumisastetta. Monet ihmiset ovat kiinnostuneita terveydestään, ja siten terveystutkimusten vastausosuudet ovat keskimäärin korkeampia kuin muissa kyselytutkimuksissa (Tourangeau et al., 2000; Groves ja Couper, 1998). Tästä huolimatta tutkimuksen molemmat vastausosuudet ovat erittäin korkeita verrattuna vastaaviin kansallisiin ja kansainvälisiin tutkimuksiin. Ne kuvastavat kenttätyön onnistumista ja aineiston edustavuutta, jota on pyritty parantamaan entisestään jälkipainotuksen eli kalibroitujen painokertoimien avulla (ks. luku 5).

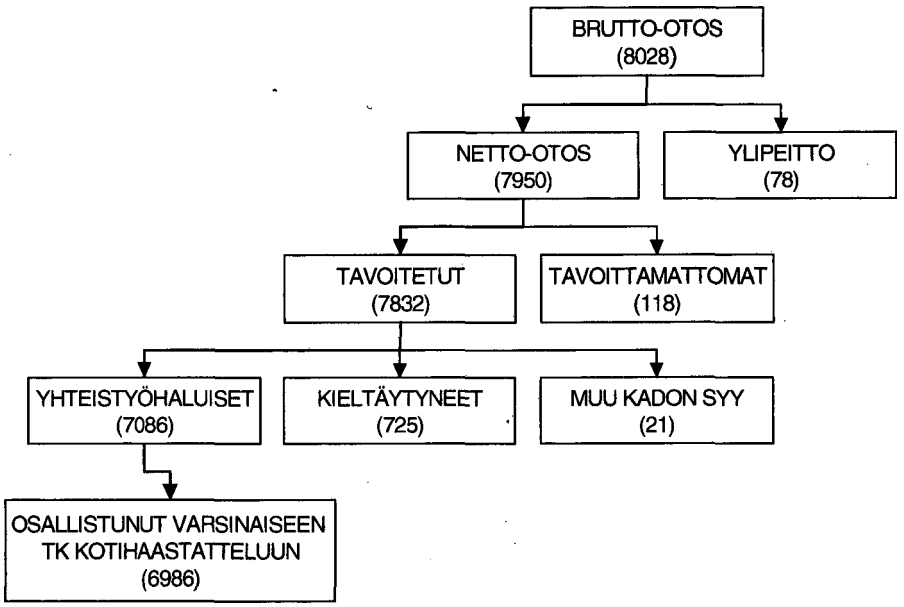
Jotta kadon oikea luonne hahmottuisi, Groves ja Couper (1998) painottavat tavallisen vastausosuuden lisäksi muiden osallistumista kuvaavien suhdelukujen tarkastelun tärkeyttä. Tällaisia suhdelukuja ovat muun muassa tavoitettujen ja kieltäytyneiden osuus netto-otokseen kuuluvista kohdehenkilöistä. Vertailukelpoisuuden vuoksi suhdeluvut tulee laskea noudattaen yleisiä standardeja (AAPOR, 2000). Koponen ja Aromaa (2003) ovat tarkastelleet tarkemmin näiden suhdelukujen käyttöä terveystutkimuksissa.

TK:n ja KTL:n haastattelijoiden sinnikkyyden, jäljittämistaitojen ja lukuisten tapaamisyritysten ansiosta tavoitettujen osuus on Terveys 2000 -tutkimuksessa erittäin korkea, peräti 98,5 prosenttia, netto-otoksesta. Myös alun perin kieltäytyneiden suostutteluun panostettiin, ja lopulta kieltäytyneiden osuus jäi 9,1 prosenttiin. Muiden syiden vuoksi katoon kuului 0,3 prosenttia otoksesta. Muita kadon syitä ovat kielivaikeudet vastata haastatteluun suomen tai ruotsin kielellä, sijaisvastaajan puute tai muu luokittelematon syy. Jos tarkastellaan painotettuja vastausosuuksia, niin tavoittamattomat henkilöt edustivat 1,5 prosenttia kohdeväestöstä (naiset 0,8 ja miehet 2,4), kieltäytyneet 8,9 prosenttia (naiset 8,4 ja miehet 9,6) ja muun syyn vuoksi katoon jäi 0,2 prosenttia (ks. kuviot 4.1 ja 4.2 ja taulukko 4.1.).

Tutkimuksessa käytetty lopputuloskoodisto kadolle on esitetty taulukossa 4.3. Koodisto poikkeaa hieman katokoodistoille ehdotetuista kansainvälisistä standardeista (ks. esim. AAPOR, 2000; Lynn et al., 2002), koska se pyrittiin räätälöimään terveystutkimukselle tarkoituksenmukaiseksi ja sopivaksi Suomen tilastotuotannon ympäristöön. Käytetty koodisto noudattaa kuitenkin ylätasolla yleisiä standardeja. Kadon syyt luokitellaan usein karkeasti tavoittamattomiin, kieltäytyneisiin ja muihin syihin (Kish, 1987; Platek and Gray, 1986; Kviz,

1977). Standardoidun katokoodiston hyödyntäminen eri tutkimuksissa mahdollistaisi tutkimukseen osallistumisen ja kadon yksityiskohtaiset vertailut.

Kuvio 4.1.
Terveys 2000 -terveyshaastatteluaineiston muodostuminen brutto-otoksesta



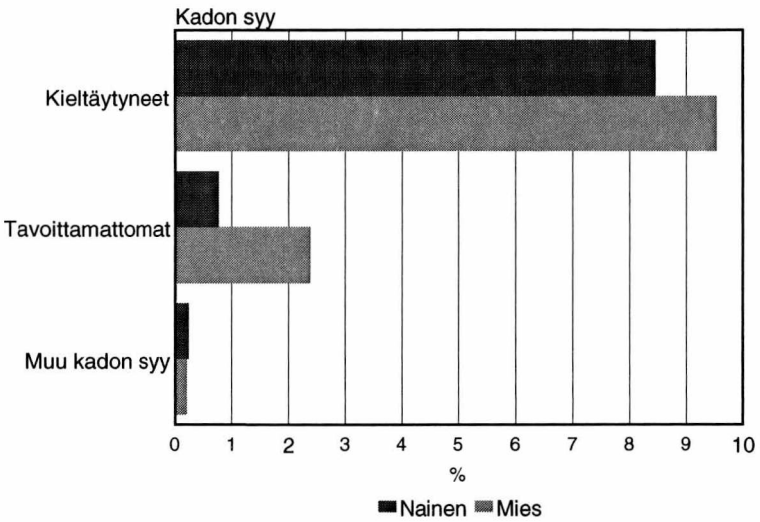
Kuviossa TK-lyhenteellä viitataan Tilastokeskukseen.

Taulukko 4.1.
Terveyshaastatteluaineiston vastauskato kadon syyn mukaan

	Katotapaukset netto-otoksessa						Kadon edustavuus perusjoukossa					
	Naisia	% -osuus	Miehiä	% -osuus	Kaikki	% -osuus	Naisia	% -osuus	Miehiä	% -osuus	Kaikki	% -osuus
Tavoittamattomat	33	0,8	85	2,4	118	1,5	13 264	0,8	35 742	2,4	49 006	1,5
– Ei tavoitettu, osoite ja puhelinnumero tiedossa	18	0,4	47	1,3	65	0,8	7 366	0,4	19 815	1,3	27 181	0,8
– Osoite tuntematon, asuinpaikkaa ei löydetä	4	0,1	21	0,6	25	0,3	1 188	0,1	8 532	0,6	9 720	0,3
– Tilapäisesti poissa	7	0,2	15	0,4	22	0,3	2 995	0,2	6 409	0,4	9 403	0,3
– Muuttanut pysyvästi muualle	2	0,0	1	0,0	3	0,0	803	0,0	531	0,0	1 333	0,0
– Tilapäisesti laitoksessa, jossa haastattelua ei voitu tehdä	2	0,0	1	0,0	3	0,0	914	0,1	455	0,0	1 368	0,0
Haastattelusta kieltäytyminen	379	8,7	346	9,6	725	9,1	146 477	8,4	143 458	9,5	289 934	8,9
– Ajanpuutteen vuoksi	99	2,3	92	2,6	191	2,4	38 801	2,2	38 662	2,6	77 463	2,4
– Periaatteellisista syistä	82	1,9	66	1,8	148	1,9	32 663	1,9	28 483	1,9	61 146	1,9
– Sairauden vuoksi	50	1,1	34	0,9	84	1,1	17 351	1,0	13 373	0,9	30 724	0,9
– Hyvän terveydentilan vuoksi*	8	0,2	14	0,4	22	0,3	3 268	0,2	5 730	0,4	8 999	0,3
– Muu erittelemätön kieltäytymisen syy	140	3,2	140	3,9	280	3,5	54 393	3,1	57 209	3,8	111 602	3,4
Muu kadon syy	13	0,3	8	0,2	21	0,3	4 126	0,2	3 276	0,2	7 402	0,2
– Sairauden tai vamman takia ei voida haastatella, ei sijaisvastaajaa	6	0,1	–	–	6	0,1	1 420	0,1	–	–	1 420	0,0
– Haastattelua ei voida tehdä kielen vuoksi	2	0,0	1	0,0	3	0,0	768	0,0	405	0,0	1 173	0,0
– Muu erittelemätön syy, ei pystytä luokittelemaan	5	0,1	7	0,2	12	0,2	1 938	0,1	2 871	0,2	4 809	0,1
Nettokato yhteensä	**425**	**9,8**	**439**	**10,1**	**864**	**10,9**	**163 866**	**9,4**	**182 475**	**12,0**	**346 342**	**10,6**

*Osa kohdehenkilöistä kieltäytyi osallistumaan suostuttelusta huolimatta. He ilmoittivat syyksi hyvän terveydentilansa, vaikka heitä taivuteltiin vetoamalla heidän osallistumisensa auttavan Terveys 2000 -tutkimusta, koska se kohdistettiin nimenomaan kaikkiin Suomessa vakituisesti asuviin henkilöihin.

Kuvio 4.2.
Kadon jakautuminen syyn mukaan



Korkeat vastausosuudet saavutettiin pienentämällä määrätietoisesti katoa jo ennen varsinaista kenttätyövaihetta. Haastattelijoita koulutettiin, ja tutkimuksesta tiedotettiin kansallisten ja paikallisten tiedotusvälineiden kautta. Lisäksi kohdehenkilöille jaettiin Terveys 2000 -tutkimuksen esitteitä.

Tutkimuksen osallistumisaste kuvaa hyvin onnistunutta haastattelijoiden kenttätyötä. TK:n haastattelijoiden kenttätyön päätyttyä tutkimuksen terveyshaastatteluosuuden vastausosuus oli peräti 87,9 prosenttia. Vain 1,7 prosenttia vastauskadosta oli tavoittamattomia kohdehenkilöitä, tutkimuksesta kieltäytyi 10,1 prosenttia ja 0,4 prosenttia ei muista syistä pystynyt osallistumaan tutkimukseen.

KTL:n haastattelijat lisäsivät entisestään TK:n haastattelijoiden saavuttamaa korkeata vastausosuutta jatkamalla kenttätyötä. Lisäyhteydenottoyrityksiä kohdennettiin 136 aiemmin tavoittamattomille kohdehenkilöille, joista tavoitettiin 18 henkilöä. Myös 799 tutkimuksesta aiemmin kieltäytyneestä henkilöstä 75 saatiin suostuteltua osallistumaan tutkimukseen. Sairauden tai tulkin puutteen vuoksi aiemmin vastaamatta jääneistä 7 henkilöä pystyi osallistumaan tutkimukseen myöhemmin. KTL:n haastattelijoiden lisätyön jälkeen tutkimukseen saatiin osallistumaan 100 henkilöä enemmän ja nettokato pieneni 12,1 prosentista 10,9 prosenttiin. Taulukossa 4.2. on esitelty kenttätyövaiheen jatkamisen vaikutusta kadon pienenemiseen.

Taulukko 4.2.
Terveyshaastatteluaineiston kadon pienentyminen kenttätyövaihetta jatkamalla

	Katotapaukset otoksessa			
	Kato TK:n kenttätyön jälkeen*	%-netto-otoksesta	Kato KTL:n kenttätyön jälkeen*	%-netto-otoksesta
Tavoittamattomat	**136**	**1,7**	**118**	**1,5**
– Kohdetta ei tavoitettu, osoite ja puhelinnumero tiedossa	76	1,0	65	0,8
– Osoite tuntematon, asuinpaikkaa ei löydetä	26	0,3	25	0,3
– Kohde tilapäisesti poissa	25	0,3	22	0,3
– Kohde muuttanut pysyvästi muualle	3	0,0	3	0,0
– Kohde tilapäisesti laitoksessa, jossa haastattelua ei voitu tehdä	6	0,1	3	0,0
Haastattelusta kieltäytyminen	**799**	**10,1**	**724**	**9,1**
– Ajanpuutteen vuoksi	200	2,5	191	2,4
– Periaatteellisista syistä	156	2,0	148	1,9
– Sairauden vuoksi	105	1,3	84	1,1
– Hyvän terveyden tilan vuoksi	24	0,3	22	0,3
– Muu erittelemätön kieltäytymisen syy	314	3,9	279	3,5
Muu kadon syy	**28**	**0,4**	**21**	**0,3**
– Sairauden tai vamman takia ei voida haastatella, ei sijaisvastaajaa	12	0,2	6	0,1
– Haastattelua ei voida tehdä kielen vuoksi	4	0,1	3	0,0
– Muu erittelemätön syy, ei pystytä luokittelemaan	12	0,2	12	0,2
Nettokato yhteensä	**963**	**12,1**	**863**	**10,9**

* TK- ja KTL-lyhenteillä viitataan Tilastokeskukseen ja Kansanterveyslaitokseen

4.2 Alueellisten erojen tarkastelu

Alueluokituksena on tässä selvityksessä käytetty NUTS-luokituksen sijaan terveystutkimukselle tarkoituksenmukaisempaa jakoa yliopistollisiin keskussairaalapiireihin (miljoonapiiri) ja terveyskeskuspiireihin, joita on myös käytetty Terveys 2000 -tutkimuksen otoksen ryvästyksessä.

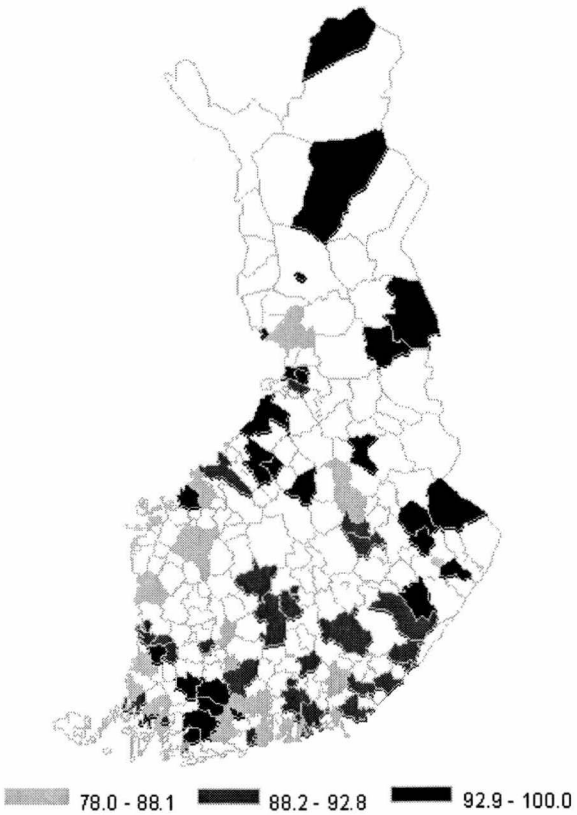
Kato-osuuksien (eli kääntäen vastausosuuksien) alueellinen vaihtelu on merkittävää Terveys 2000 -tutkimuksessa. Karttakuviosta 4.3. nähdään, että vastausosuudet olivat alimmillaan pääkaupunkiseudulla, muissa Etelä- ja Keski-Suomen kaupungeissa ja niiden läheisissä terveyskeskuspiireissä ja korkeimmillaan haja-asutusalueilla ja Pohjois-Suomessa.

Kuviossa 4.4. on esitetty tutkimuksen painotettu kato-osuus miljoonapiirin ja sukupuolen mukaan. Kohdehenkilöiden kato-osuus on keskimääräistä korkeampi Helsingin ja Tampereen miljoonapiirissä. Naisten kato-osuudet ovat alhai-

simmillaan Turun, Kuopion ja Oulun miljoonapiireissä. Sen sijaan miesten vastausaktiivisuus on lähellä maan keskitasoa Turun ja Kuopion miljoonapiireissä. Miehet osallistuivat tutkimukseen naisiakin aktiivisemmin Oulun miljoonapiirissä.

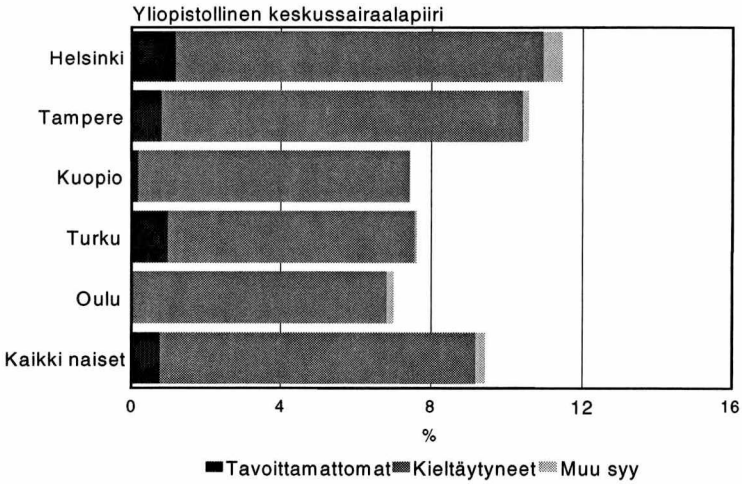
Oulun miljoonapiirissä haastattelijat onnistuivat erittäin hyvin tavoittamaan tutkimukseen poimitut henkilöt. Muualla Suomessa miesten tavoittaminen oli vaikeampaa kuin naisten. Erityisesti Helsingin ja Turun miljoonapiireissä tavoittamattomuus jäi miesten merkittäväksi kadon syyksi. Tavoittamattomuus oli hyvin alhaista Terveys 2000 -tutkimuksessa. Yleisin kadon syy oli tutkimuksesta kieltäytyminen, joka oli ongelmallisinta Helsingin ja Tampereen miljoonapiireissä.

Kuvio 4.3.
Painottamaton vastausosuus terveyskeskuspiireittäin

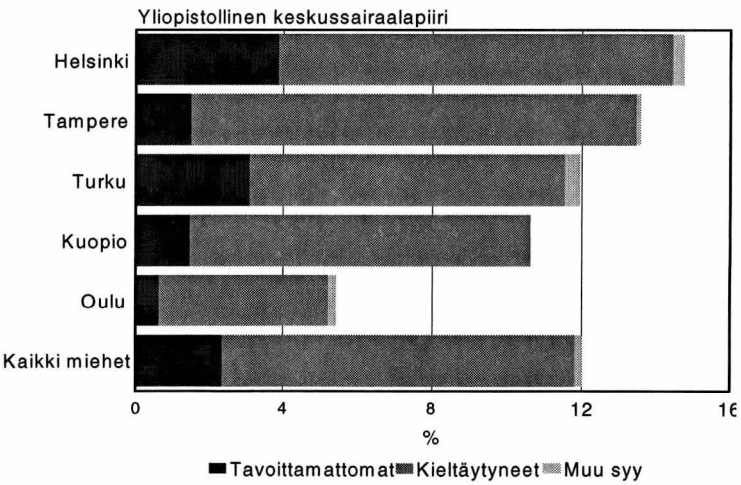


Kuvio 4.4.
Kato-osuus yliopistollisen keskussairaalapiirin ja sukupuolen mukaan

Naiset:



Miehet:



4.3 Kato eri väestöryhmissä

Katoanalyysissa lopullisen aineiston rakennetta verrataan tutkimuksen perusjoukon kanssa. Vastauskatoa on perinteisesti tarkasteltu kohdehenkilöiden tai heidän asuntokuntien sosiodemografisten tekijöiden mukaan. Jakaumien vertailun mahdollistavat eri rekisterit ja hallinnolliset aineistot, joista on voitu yhdistää yksilökohtaista tietoa kaikille otokseen poimituille. Tässä luvussa tarkastelut on esitetty kadon syyn mukaan.

Groves ja Couper (1998) ovat erottaneet kohteiden tavoitettavuuden ja tavoitettujen osallistumiskäyttäytymisen analysoinnin. He ovat kehittäneet mallin kohteen tavoittamisesta sekä surveytutkimusten osallistumisen käsitteellisen kehikon. Kohteen tavoittamisen onnistumiseen vaikuttavat tavoittamisyritysten määrän ja ajoituksen lisäksi sosiaalisen ympäristön tekijät, sosiodemografiset piirteet, haastattelijan kohtaamat fyysiset lähestymisesteet kohteen asuinpaikassa sekä kohteen ajankäyttö kotona (tai tavoitettavuus puhelimitse).

Surveytutkimusten osallistumisen käsitteellisessä kehikossa tarkastellaan tavoitetun kohdehenkilön/-kotitalouden suostumista osallistua tutkimukseen. Kohde ja haastattelija toimivat samassa sosiaalisessa ympäristössä ja tutkimusasetelman vaikutusten alla. Ympäristötekijät sekä henkilökohtaiset tai kotitalouden ominaisuudet voivat vaikuttaa haastattelijan ja haastateltavan vuorovaikutukseen, minkä perusteella tavoitettu kohde tekee päätöksen tutkimukseen osallistumisesta.

Määrittäessään kriittisiä Groves ja Couper (1998) viittaavat aiempiin tutkimuksiin, joissa muun muassa sukupuolen, iän, kotitalouksien koon, tulojen ja sosioekonomisen aseman on raportoitu vaikuttavan osallistumisasteeseen (ks. esim. Glenn, 1969; Kemsley, 1975; Smith, 1983; Lindström, 1983; Ekholm ja Laaksonen, 1991). Kuvaileva katoanalyysi on tehty pääasiallisesti sukupuolen mukaan, koska miesten ja naisten vastauskäyttäytyminen voi poiketa toisistaan eri taustatekijöiden mukaan.

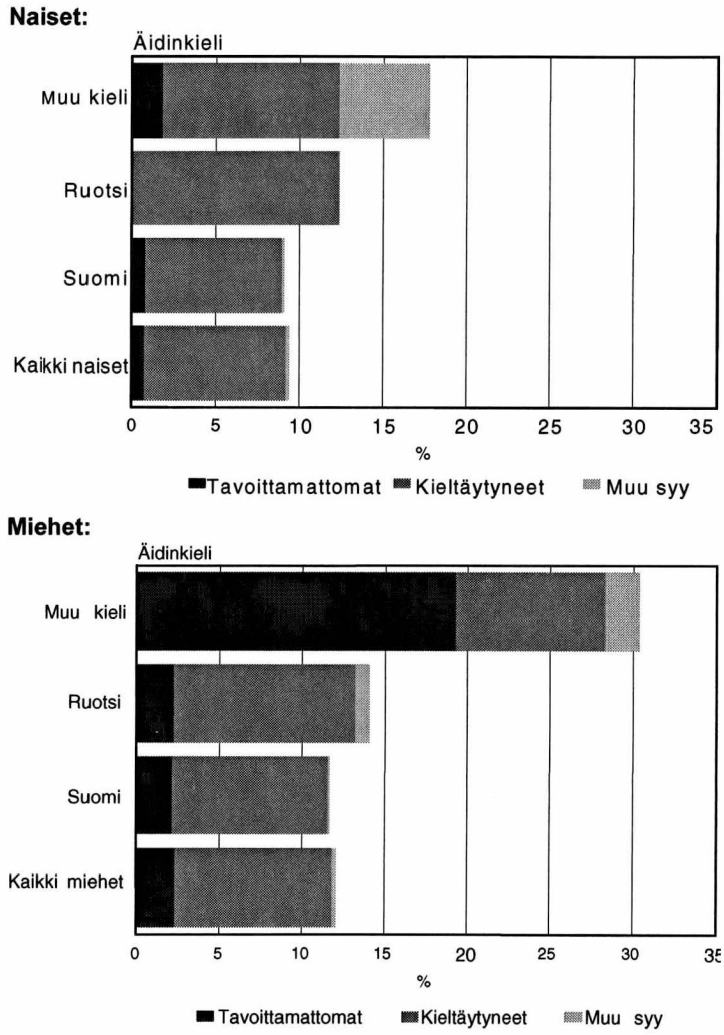
Äidinkieli

Äidinkielitiedot perustuvat väestötietojärjestelmään. Käytetyssä kieliluokituksessa suomen ja saamen kieltä puhuvat on yhdistetty samaan luokkaan. Toisen luokan muodostuvat ruotsin kielen äidinkielekseen rekisteröineet; muita kieliä äidinkielenään puhuvat on luokiteltu vastaavasti omaksi luokakseen.

Terveyshaastattelu tehtiin joko suomen tai ruotsin kielellä. Kato-osuus on suurin muuta kuin kotimaisia kieliä äidinkielenään puhuvien keskuudessa: naisilla 18 prosenttia ja miehillä peräti 31 prosenttia. Myös ruotsinkielisten osallistuminen tutkimukseen oli hieman vähäisempää kuin suomenkielisten.

Tavoitettavuus on yleisin kadon syy muuta kieltä puhuvien miesten keskuudessa. Kaikissa muissa väestöryhmissä yleisin syy on tutkimuksesta kieltäytyminen. Vierasta kieltä äidinkielenään puhuvat jäivät keskimääräistä useammin kadoksi nimenomaan kielivaikeuksien vuoksi. Koska terveydentilalla ja äidinkielellä voi olla yhteyksiä, kadon vaikutusta on pyritty pienentämään kalibroimalla alkuperäiset painokertoimet muuan muassa kohdehenkilöiden äidinkielen mukaan (ks. luku 5).

Kuvio 4.5.
Kato-osuus äidinkielen ja sukupuolen mukaan

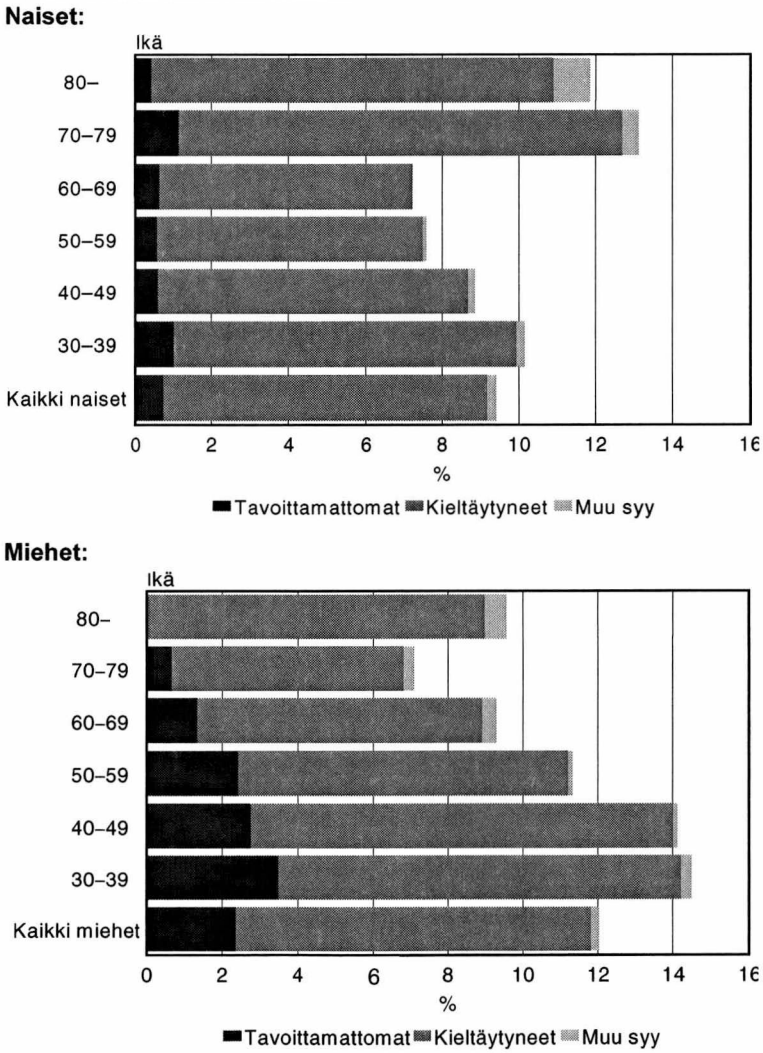


Ikä

Vastauskäyttäytymisessä on merkittäviä eroja henkilöiden iän ja sukupuolen mukaan (Kuvio 4.6.). Iäkkäiden naisten ja nuorempien miesten todennäköisyys osallistua tutkimukseen oli alhaisin. Yli 70-vuotiaiden naisten kato-osuus oli merkittävästi korkeampi kuin keskimäärin. Myös nuorten 30–39-vuotiaiden naisten vastausaktiivisuus oli hieman keskimääristä alhaisempi, mutta selkeästi vanhimpia ikäluokkia korkeampi. Miesten kato-osuus oli suuri 30–49-vuotiailla. Sitä vastoin yli 50-vuotiaiden miesten vastausaktiivisuus oli jokaisessa ikäluokassa miesten keskitasoa korkeampi.

Miesten iällä ja tavoittamattomien osuudella on merkittävä yhteys. Nuorimpia miehiä oli vaikea tavoittaa haastattelijoiden useista yhteydenottoyrityksistä huolimatta. Naisten ikäryhmittäisessä tarkastelussa ei ilmene merkittäviä eroja iän ja kadon syyn mukaan.

Kuvio 4.6.
Kato-osuus iän ja sukupuolen mukaan



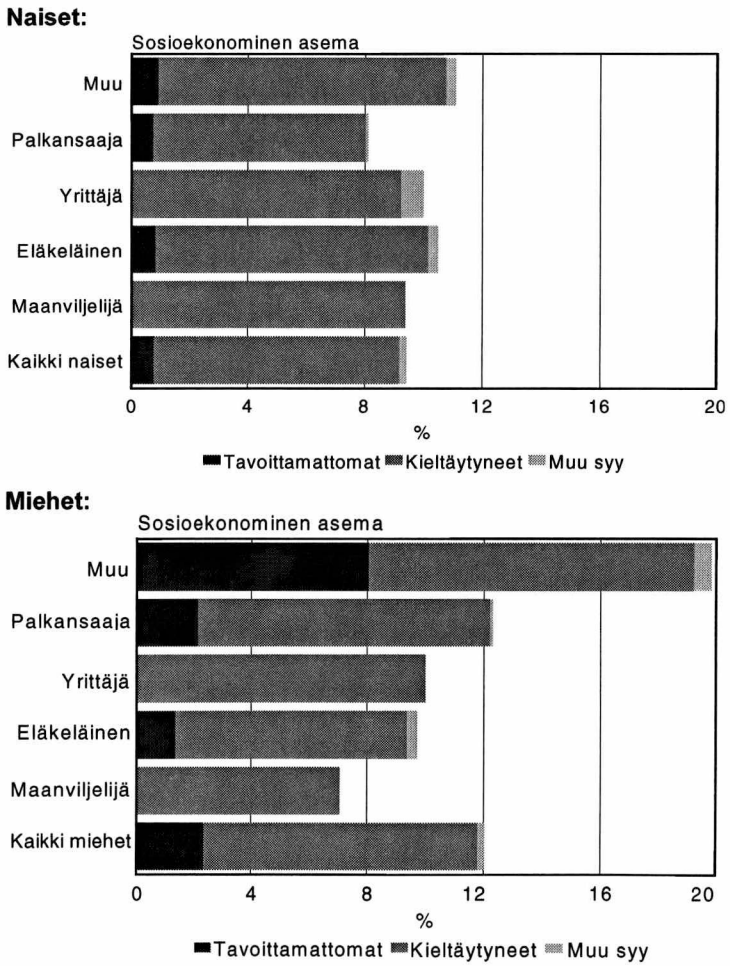
Sosioekonominen asema

Tarkasteltaessa vastausosuuksia rekistereistä johdetun karkean sosioekonomisen aseman luokituksen ja sukupuolen mukaan havaitaan enemmän kato-osuuksien hajontaa erityisesti miesten kohdalla (Kuvio 4.7.). Niin kutsuttuun residuaaliryhmään 'Muu' kuuluvat ne, joita ei voida luokitella mihinkään muuhun so-

sioekonomiseen ryhmään. Heistä erityisesti miehet osallistuivat muita harvemmin tutkimukseen.

Kaikki nais- ja miespuoliset yrittäjät sekä maanviljelijät tavoitettiin tutkimuksessa; maanviljelijöiden ja miespuolisten yrittäjien ainoa kadon syy on kieltäytyminen. Tavoittamattomien osuus oli suurimmillaan niiden miesten keskuudessa, jotka luokiteltiin pääasiallisen sosioekonomisen aseman "muuhun" ryhmään.

Kuvio 4.7.
Kato-osuus sosioekonomisen aseman ja sukupuolen mukaan

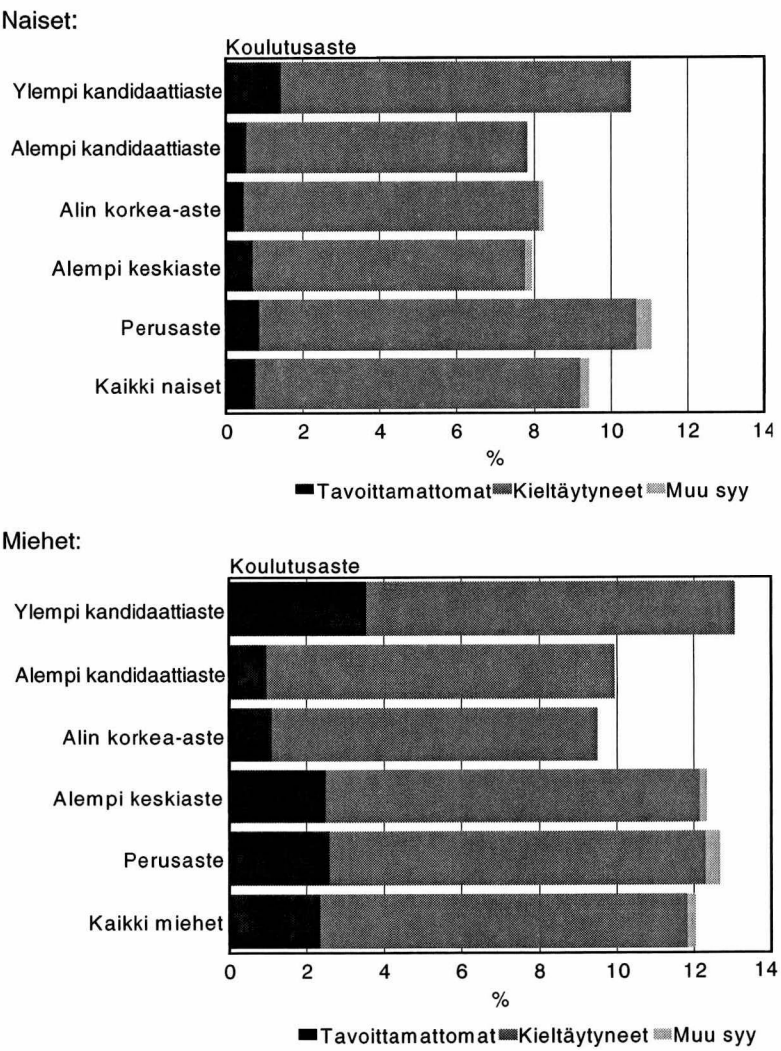


Koulutusaste

Koulutustaustan mukaan tarkasteltuna nähdään, että kato on suurin ylemmän kandidaattiasteen tai tutkijakoulutuksen suorittaneilla ja vain perusasteen koulutuksen omaavilla (Kuvio 4.8.). Tavoittamattomien osuus on suurin ylemmän kandidaattiasteen suorittaneilla kuin muun koulutustaustan omaavilla. Kadon

muut syyt, kuten esimerkiksi kielivaikeudet keskittyvät alhaisemman koulutustaustojen pariin.

Kuvio 4.8.
Kato-osuus koulutuksen ja sukupuolen mukaan



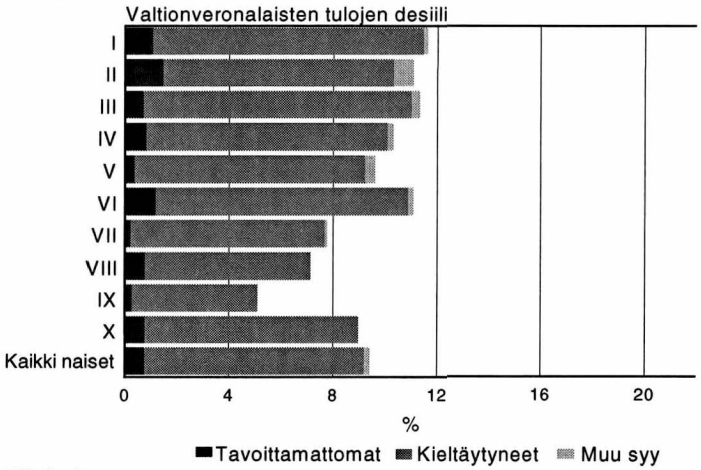
Tulot

Kuviossa 4.9. on tarkasteltu tutkimukseen osallistumista valtion veronalaisten tulojen mukaan. Koko väestölle on laskettu tulodesiilirajat, ja henkilöt on sijoitettu vastaaviin desiililuokkiin. Tutkimukseen osallistuminen vaihtelee tuloluokan mukaan. Kato-osuudet ovat keskimääräistä suurempia alemmissa tuloluokissa ja pienevät ylemmissä tuloluokissa. Suurituloisten kato-osuus nousee jälleen hieman, mutta jää edelleen alle keskimääräisen kato-osuuden. Kato-

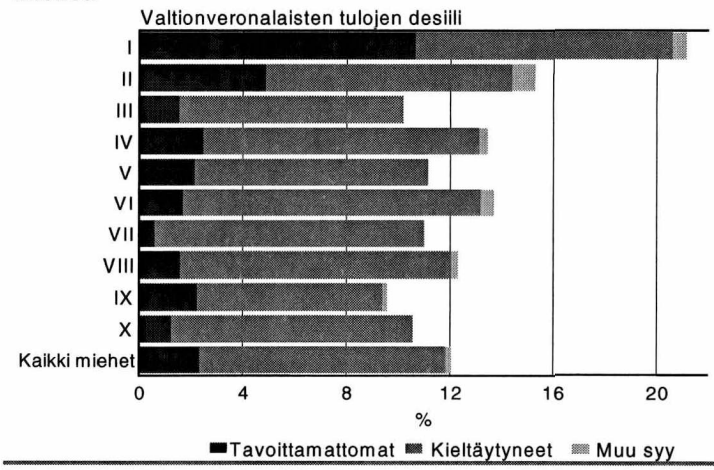
osuus on merkitsevästi suurin ensimmäisessä eli vähätuloisimpien miesten desiilissä, joista lähes kolmannes jää tutkimuksen ulkopuolelle.

Kuvio 4.9.
Kato-osuus valtionveronalaisten tulojen ja sukupuolen mukaan

Naiset:



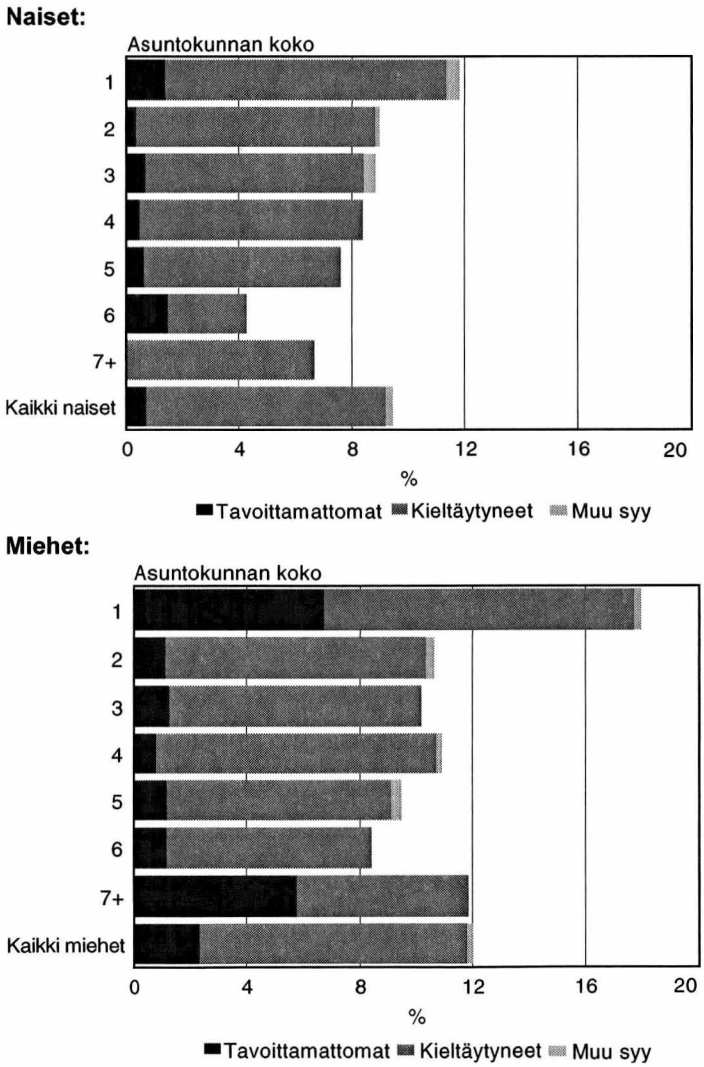
Miehet:



Asuntokunta ja perheasema

Tässä tutkimuksessa asuntokunnan koko on merkitsevä kato-osuuden selittäjä (kuvio 4.10.). Yksinasuvien naisten ja miesten kato-osuus on korkein. Suuremmissa asuntokunnissa kato on vähäistä tai lähellä keskitasoa. Erityisesti yksin tai erittäin suurissa asuntokunnissa asuvien miesten tavoitettavuus oli vaikeata.

Kuvio 4.10.
Kato-osuus asuntokunnan koon ja sukupuolen mukaan



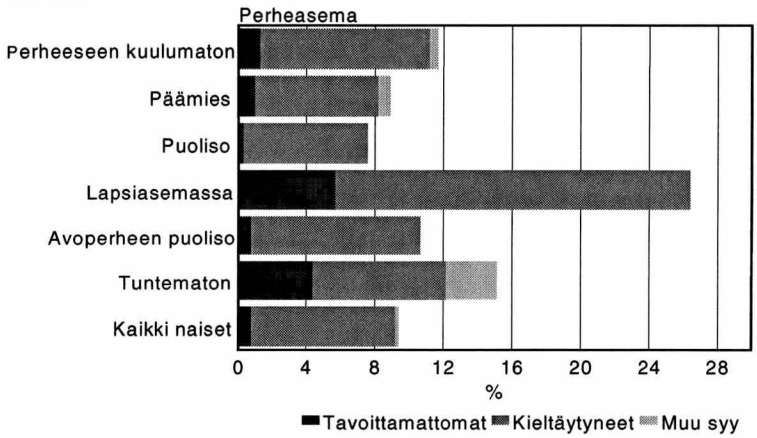
Rekistereihin perustuvassa väestötilastojen tuotannossa pystytään määrittämään asuntokohtaisten tietojen perusteella henkilöiden perheasema. Määritelmän mu-

kaan perheen muodostavat yhdessä asuvat avio- tai avoliitossa olevat henkilöt ja heidän lapsensa, toinen vanhemmista lapsineen sekä puolisot, joilla ei ole lapsia. Lapsiksi on määritelty iästä ja siviilisäädystä riippumatta vanhempiensa luona asuvat omat lapset, puolison biologiset lapset tai ottolapset. (Tilastokeskus, 2001).

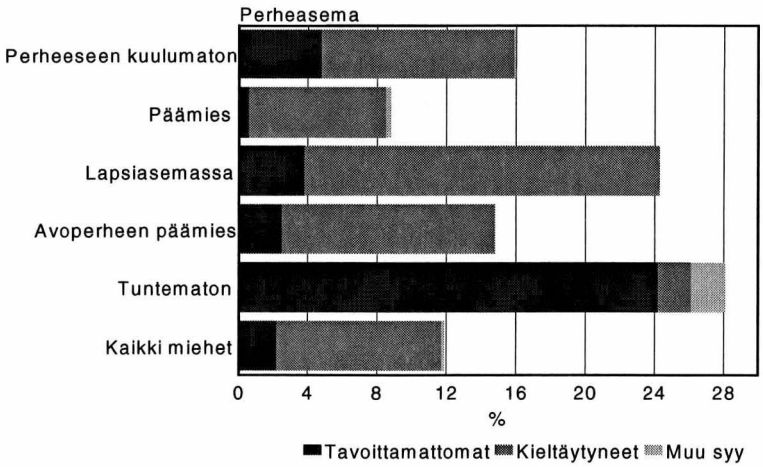
Vanhempiensa luona asuvien aikuisten naisten ja miesten kato-osuus on tutkimuksessa hyvin korkea: heistä peräti joka neljäs ei osallistunut tutkimukseen. Perheasemaltaan tuntemattomien ja perheeseen kuulumattomien kato-osuudet ovat korkeita niin naisilla kuin miehillä (Kuvio 4.11.).

Kuvio 4.11.
Kato-osuus perheaseman ja sukupuolen mukaan

Naiset:



Miehet:



4.4 Logistinen regressioanalyysi

Tavoitettavuuteen ja tutkimukseen osallistumiseen vaikuttavia tekijöitä on tutkittu netto-otoksen 30 vuotta täyttäneille henkilöille logististen regressiomallien avulla. Analyysissa on hyödynnetty eri rekistereistä, hallinnollisista aineistoista ja tilastoista saatavan lisäinformaatiota niin henkilö-, asuntokunta- kuin terveyskeskuspiiritasolla. Ensimmäinen logistinen regressiomalli kuvaa tavoittamisen todennäköisyyttä. Selitettävänä muuttujana on binäärinen vastausindikaattorimuuttuja, joka saa arvon 1, jos kohdehenkilö on tavoitettu, ja arvon 0, jos henkilö kuuluu katoon. Toinen malli on ehdollinen onnistumiselle ensimmäisessä mallissa eli kohteen tavoittamiselle eli se mallittaa tavoitettujen kohdehenkilöiden vastauskäyttäytymistä. Siinä binäärinen vastausindikaattorimuuttuja saa arvon 1, jos henkilö on vastannut hyväksyttävästi Terveys 2000 -tutkimuksen pitkään tai lyhyeen terveyshaastatteluun, ja arvon 0, jos henkilö kuuluu katoon.

Otantaan perustuva aineisto on ryvästynyt terveyskeskuspiireittäin, jotka ovat tilastollisesti merkitsevä alueellinen taso mallitettaessa tutkimukseen poimittujen henkilöiden vastaustodennäköisyyksiä (Laiho, 2001). Lisäksi molemmissa malleissa muutama terveyskeskuspiiritason muuttuja on tilastollisesti merkitsevä. Tämän vuoksi on perusteltua käyttää malliasetelmana otanta-asetelman huomioivaa logit-mallia, jonka rypäät muodostuvat otoksen 80 terveyskeskuspiiristä. Malli hyödyntää painokertoimina alkuperäisistä sisältymistodennäköisyyksistä johdettuja asetelmapainoja. Analyysi on tehty SUDAAN-ohjelmiston LOGISTIC/RLOGIST-proseduurilla.

Logistisen regressiomallin tulokset tiivistävät aiemmin luvussa esitetyn kuvailevan yhden muuttujan analyysin. Se myös mahdollistaa saman aikaisen vertailun eri tekijöiden merkittävyydestä vastauskadon selittäjinä. Tärkeimmiksi selittäviksi muuttujiksi tässä analyysissa osoittautuneet muuttujat ja niiden diagnostiikka on esitetty taulukossa 4.3. kohdehenkilöiden tavoitettavuudelle ja taulukossa 4.4 terveyshaastatteluun osallistumiselle ehdolla, että kohdehenkilöt tavoitettiin.

Tavoitettavuutta vähentää kohdehenkilön asumis- ja perhetilannetta kuvaavat muuttujat: perheettömyys tai yksinasuminen, aikuisena henkilönä vanhempien luona asuminen (lapsiasemassa) ja kotitalousväestöön kuulumattomuuteen johtaneet syyt, kuten asunnottomuus tai laitosväestöön kuuluminen. Lisäksi kohdehenkilöiden tavoitettavuus on vähän vaikeampaa terveyskeskuspiireissä, joissa rikollisuusaste on keskimääräistä korkeampi. Vastaavasti yrittäjien osuus lisää tavoitettujen osuutta terveyskeskuspiireissä. Myös korkea tulotaso ja pääomatulojen saanti lisäävät tavoittamisen todennäköisyyttä. Eri tekijöiden yhteisvaikutukset eivät olleet tavoitettavuuden mallissa merkitseviä (taulukko 4.3.).

Taulukko 4.3.
Henkilöiden tavoitettavuuteen vaikuttavat tekijät logit-mallissa

Muuttujat	Beta-estimaatti	Keski-virhe	t-testi	p-arvo
Vakio	2,62	0,70	3,7	0,000
Ikä	0,02	0,01	3,8	0,000
Sukupuoli (1=nainen, 2=mies)				
1	-1,11	0,21	-5,2	0,000
2	0,00	0,00	.	.
Muu äidinkieli kuin suomi, ruotsi tai saame (1=kyllä, 2=ei)				
1	-1,44	0,38	-3,8	0,000
2	0,00	0,00	.	.
Perheetön (1=kyllä, 2=ei)				
1	-1,27	0,19	-6,6	0,000
2	0,00	0,00	.	.
Lapsiasemassa (1=kyllä, 2=ei)				
1	-1,32	0,44	-3,0	0,003
2	0,00	0,00	.	.
Ln(Valtion verotuksessa veronalaiset ansiotulot)	0,15	0,03	5,0	0,000
Pääomatuloja (1=kyllä, 2=ei)				
1	0,87	0,28	3,1	0,002
2	0,00	0,00	.	.
Kuuluu kotitalousväestöön (1=ei, 2=kyllä)				
1	-2,25	0,30	-7,5	0,000
2	0,00	0,00	.	.
Keskimääräinen rikosten lkm per asukas TKP:ssä	-6,85	2,61	-2,6	0,009
Yrittäjien %-osuus TKP:ssä	0,14	0,04	4,1	0,000
Oulun miljoonapiiri (1=kyllä, 2=ei)				
1	1,94	0,72	2,7	0,007
2	0,00	0,00	.	.

Muuttuja	OR[1)]	OR:n 95% luottamusväli alaraja	yläraja
Vakio	13,76	3,49	54,34
Ikä	1,02	1,01	1,03
Sukupuoli (1=nainen, 2=mies)			
1	0,33	0,22	0,50
2	1,00	1,00	1,00
Muu äidinkieli kuin suomi, ruotsi tai saame (1=kyllä, 2=ei)			
1	0,24	0,11	0,50
2	1,00	1,00	1,00
Perheetön (1=kyllä, 2=ei)			
1	0,28	0,19	0,41
2	1,00	1,00	1,00
Lapsiasemassa (1=kyllä, 2=ei)			
1	0,27	0,11	0,63
2	1,00	1,00	1,00
Ln(Valtion verotuksessa veronalaiset ansiotulot)	1,17	1,10	1,24
Pääomatuloja (1=kyllä, 2=ei)			
1	2,39	1,38	4,13
2	1,00	1,00	1,00
Kuuluu kotitalousväestöön (1=ei, 2=kyllä)			
1	0,11	0,06	0,19
2	1,00	1,00	1,00
Keskimääräinen rikosten lkm per asukas TKP:ssä	0,00	0,00	0,18
Yrittäjien %-osuus TKP:ssä	1,16	1,08	1,24
Oulun miljoonapiiri (1=kyllä, 2=ei)			
1	6,95	1,69	28,64
2	1,00	1,00	1,00

[1)] OR eli odds ratio tarkoittaa suhteellista riskiä mallissa selittävälle tapahtumalle

Taulukko 4.4.
Lyhyeen tai pitkään kotihaastatteluun vastaamiseen vaikuttavat tekijät logit-mallissa

Muuttujat	Beta-estimaatti	Keski-virhe	t-testi	p-arvo
Vakio	0,07	0,71	0,1	0,924
Ikä	0,00	0,00	-0,3	0,747
Sukupuoli (1=nainen, 2=mies)				
1	-0,88	0,30	-2,9	0,004
2	0,00	0,00	.	.
Lapsiasemassa (1=kyllä, 2=ei)				
1	-0,95	0,22	-4,4	0,000
2	0,00	0,00	.	.
Valtion verotuksessa veronalaisten ansiotulojen desiili otoksessa	0,04	0,02	2,7	0,008
Asuu rivi- tai paritalossa	0,30	0,13	2,3	0,022
Saanut pääomatuloja	0,29	0,10	2,9	0,004
Saanut sairaspäivärahaa	0,44	0,19	2,3	0,022
Ln(Huoneiston pinta-ala)	0,46	0,09	4,9	0,000
Huoltosuhde TKP:ssä	0,62	0,16	3,9	0,000
15-24 vuotiaiden %-osuus TKP:ssä	-0,08	0,03	-2,6	0,010
Oulun miljoonapiiri (1=kyllä, 2=ei)	0,50	0,17	2,9	0,003
Ikä, Sukupuoli (1=nainen, 2=mies)				
1, 1	0,01	0,01	2,4	0,015
1, 2	0,00	0,00	.	.

Muuttuja	OR	OR:n 95 %luottamusväli alaraja	yläraja
Vakio	1,07	0,26	4,33
Ikä	1,00	0,99	1,01
Sukupuoli (1=nainen, 2=mies)			
1	0,41	0,23	0,75
2	1,00	1,00	1,00
Lapsiasemassa (1=kyllä, 2=ei)			
1	0,39	0,25	0,59
2	1,00	1,00	1,00
Valtion verotuksessa veronalaisten ansiotulojen desiili otoksessa	1,05	1,01	1,08
Asuu rivi- tai paritalossa	1,35	1,04	1,75
Saanut pääomatuloja	1,34	1,10	1,62
Saanut sairaspäivärahaa	1,56	1,06	2,27
Ln(Huoneiston pinta-ala)	1,58	1,31	1,90
Huoltosuhde TKP:ssä	1,86	1,37	2,55
15-24 vuotiaiden %-osuus TKP:ssä	0,92	0,87	0,98
Oulun miljoonapiiri (1=kyllä, 2=ei)	1,65	1,18	2,30
Ikä, Sukupuoli (1=nainen, 2=mies)			
1, 1	1,01	1,00	1,02
1, 2	1,00	1,00	1,00

Taulukoissa 4.3. ja 4.4. esitetyt muuttujat kuvaavat kohtuullisesti tutkimukseen poimittujen henkilöiden vastauskäyttäytymistä. Myös otanta-asetelman huomioonottava logit-malli tuottaa hieman paremman tuloksen kuin tavallinen logistinen regressiomalli. Tästä huolimatta suuri osa mallivaihtelusta jää edelleen selittämättä ja mallien tulokset voidaan tulkita vain suuntaa antaviksi.

Vastauskäyttäytymismalleja tulkittaessa tulee huomioida, että ne pystyvät vain rajatussa määrin kuvaamaan vastauskäyttäytymisen yhteyksiä eri väestöryhmiin tai muihin taustatekijöihin. Vastauskäyttäytymisen tarkastelu vain tilastollisin mallein voi kuitenkin olla rajallista, koska monet tavoittamattomuuden ja kieltäytymisen perimmäiset syyt eivät ole mitattavissa yhteismitallisesti. Tämän lisäksi laajakaan rekistereihin ja hallinnollisiin tiedostoihin perustuva lisäinformaatio ei välttämättä pysty selittämään tyhjentävästi yksilötason käyttäytymistä tai reaktiota tiettynä hetkenä.

5 Asetelmapohjainen estimointi

Kari Djerf, Johanna Laiho ja Tommi Härkänen

5.1 Painokertoimien muodostus ja käyttö

Painokertoimien tarkoituksena on palauttaa havaintoaineisto vastaamaan alkuperäisen perusjoukon jakaumia. Sillä siis voidaan korjata otannasta ja vastauskadosta aiheutuneita virheitä otanta-aineistossa. Painokertoimia tulee käyttää kaikissa otanta-aineistosta tehtävissä tilastollisissa analyyseissa ja estimoinneissa. Jos painokertoimia ei käytetä, aineistosta tuotetut tulokset eivät ole yleistettävissä koko kohdeperusjoukkoon eivätkä ne myöskään ole vertailukelpoisia muiden vastaavien tutkimusten kanssa.

Painokertoimien muodostaminen aloitetaan laskemalla alkuperäiset sisältymistodennäköisyydet netto-otokseen kuuluville otoshenkilöille. Henkilöiden sisältymistodennäköisyys riippuu väestön alueellisesta jakautumisesta sekä kohdehenkilöiden iästä seuraavasti. Ensimmäisen asteen poiminnassa muodostettiin aluerypäät: 15 suurinta terveyskeskuspiiriä poimittiin todennäköisyydellä 1, ja 65 terveyskeskuspiiriä poimittiin vaihtelevin todennäköisyyksin. Toisen asteen poiminnassa ositettiin poimitut rypäät iän mukaan, jolloin yli 80-vuotiaille asetettiin tiheämpi poimintaväli.

Yli 80-vuotiaiden poimintaväli on puolet pienempi kuin muilla henkilöillä eli heidän sisältymistodennäköisyytensä on kaksinkertainen samalla alueella asuviin nuorempiin henkilöihin nähden. Painokertoimien perusmuoto on siis asetelmapaino, joka useimpien otanta-asetelmien tapauksessa muodostuu kunkin kohdehenkilön sisältymistodennäköisyyden käänteislukuna. Kunkin poimitun henkilön todennäköisyys tulla valituksi on sen tähden johdettu 1 ja 2 asteen poimintasääntöjen ja väestöjakaumien perusteella.

Asetelmapaino ei useimmissa käytännön tilanteissa riitä, koska otoksen poiminnan ja tiedonkeruun jälkeen aineisto saatetaan havaita vinoksi kehikkovirheiden, otantavirheen, kadon tai mittausvirheiden vuoksi. Asetelmapainoja joudutaan muokkaamaan erilaisten mallioletusten perusteella, ja näin johdettua painotusta kutsutaan uudelleenpainotukseksi. Se on yleisesti käytössä oleva menetelmä, jolla alkuperäisiä asetelmapainoja muokataan käyttämällä hyödyksi lisäinformaatiota joko perusjoukosta, otoksesta tai kummastakin (Oh ja Scheuren, 1983). Yksinkertaisimmat uudelleenpainotusmenetelmät ovat jälkiositus sekä suhdetehostus. Jälkiosituksessa otos painotetaan perusjoukon tunnettujen jakaumatietojen mukaan, esimerkiksi henkilöpohjaisissa tutkimuksissa demografisten tietojen, kuten ikä- ja sukupuoliryhmien sekä asuinalueen mukaan (Särndal et al., 1992; Djerf, 2000).

Alkuperäisten asetelmapainojen kalibroinnilla on kaksi painokertoimia korjaavaa vaikutusta:

- korjataan kadon vaikutus lopulliseen saavutettuun otokseen ja
- yleistetään lopullinen aineisto edustamaan tutkimuksen kohdeperusjoukkoa.

Väestöjakaumat perusjoukossa ja Terveys 2000 -tutkimuksen otoksessa

Väestöjakaumat on laskettu 30 vuotta täyttäneille otanta-asetelman väestörajausta noudattaen. Tämän vuoksi Ahvenanmaata, ulkosaaristokuntia sekä niin kutsutuista 900-ryhmistä* ulkomailla tilapäisesti asuvia ei ole laskettu mukaan kohdeperusjoukkoon.

Taulukko 5.1.
Väestöjakauma kohdeperusjoukossa (30 vuotta täyttäneet) ja lopullisessa otoksessa sukupuolen ja ikäluokan mukaan

	Kohdeperusjoukko				Terveyshaastatellut otoshenkilöt			
Ikä	**Naiset**	**%**	**Miehet**	**%**	**Naiset**	**%**	**Miehet**	**%**
30–39	351 929	16,9	366 380	18,9	761	19,4	694	21,9
40–49	383 007	18,4	392 888	20,3	854	21,8	772	24,4
50–59	365 789	17,6	367 322	19,0	759	19,3	748	23,7
60–69	254 092	12,2	225 395	11,7	548	14,0	470	14,9
70–79	226 153	10,9	146 129	7,6	462	11,8	294	9,3
80+ [1]	127 809	6,2	47 788	3,0	540	13,8	184	5,8
Yhteensä	**1 708 779**	**100,0**	**1 545 902**	**100,0**	**3 924**	**100,0**	**3 162**	**100,0**

1) Otanta-asetelmassa 80 vuotta täyttäneiden sisältymistodennäköisyys määriteltiin kaksinkertaiseksi

Kohdeperusjoukon väestöjakaumat on johdettu väestötietojärjestelmästä. Taulukossa 5.2. on esitetty väestön määrä todennäköisyydellä 1 poimituissa terveyskeskuspiireissä ja jäljelle jäävissä miljoonapiireissä eli Helsingin (HYKS), Turun (TYS), Tampereen (TAYS), Kuopion (KYS) ja Oulun (OYS) yliopistollisissa keskussairaalapiireissä.

* Väestöjakaumat on laskettu 30 vuotta täyttäneille otanta-asetelman väestörajausta noudattaen. Tämän vuoksi Ahvenanmaata, ulkosaaristokuntia sekä niin kutsutuista 900-ryhmistä ulkomailla tilapäisesti asuvia ei ole laskettu mukaan kohdeperusjoukkoon.

Väestörekisterissä 900-ryhmään kuuluvat mm. laitosväestö, vailla vakituista asuntoa olevat sekä tilapäisesti ulkomaille muuttaneet. Ulkomailla asuva, mutta Suomessa kirjoilla oleva väestö rajattiin Terveys 2000 -tutkimuksen perusjoukon ulkopuolelle.

Taulukko 5.2.
Väestöjakauma kohdeperusjoukossa (30 vuotta täyttäneet) ja lopullisessa otoksessa suurten terveyskeskuspiirien ja jäljelle jäävien miljoonapiirien mukaan

	Väestö	%	Terveys-haastattelu[1]	%	Terveys-tarkastus[2]	%	Terveys-kyselyt[3]	%	Terveys 2000 -tutkimus[4]	%
Espoo	123 988	3,8	269	3,8	259	3,8	262	3,8	269	3,8
Helsinki	346 470	10,7	743	10,5	697	10,3	705	10,3	748	10,5
Vantaa	105 999	3,3	225	3,2	214	3,2	215	3,2	226	3,2
Kotka	36 794	1,1	77	1,1	72	1,1	72	1,1	77	1,1
Lappeen-ranta	36 957	1,1	78	1,1	78	1,2	77	1,1	79	1,1
Turku	107 280	3,3	241	3,4	230	3,4	233	3,4	243	3,4
Pori	49 957	1,5	110	1,6	106	1,6	106	1,6	110	1,6
Hämeen-linna	45 144	1,4	90	1,3	85	1,3	86	1,3	91	1,3
Tampere	119 856	3,7	268	3,8	258	3,8	259	3,8	269	3,8
Lahti	62 663	1,9	130	1,8	131	1,9	131	1,9	131	1,8
Vaasa	34 289	1,1	72	1,0	66	1,0	66	1,0	73	1,0
Joensuu	30 882	1,0	64	0,9	63	0,9	63	0,9	65	0,9
Kuopio	52 852	1,6	115	1,6	109	1,6	110	1,6	115	1,6
Jyväskylä	45 470	1,4	101	1,4	94	1,4	96	1,4	101	1,4
Oulu	67 303	2,1	166	2,3	162	2,4	162	2,4	167	2,4
Muu HYKS	411 965	12,7	853	12,0	809	12,0	818	12,0	857	12,1
Muu TYKS	281 717	8,7	640	9,0	611	9,0	612	9,0	643	9,0
Muu TAYS	495 642	15,2	1 041	14,7	1 003	14,8	1 012	14,8	1 044	14,7
Muu KYS	429 846	13,2	965	13,6	919	13,6	930	13,6	965	13,6
Muu OYS	369 607	11,4	838	11,8	804	11,9	813	11,9	839	11,8
Yhteensä	**3 254 681**	**100,0**	**7 086**	**100,0**	**6 770**	**100,0**	**6 828**	**100,0**	**7 112**	**100,0**

[1] Terveyshaastatteluun osallistuneet

[2] Terveystarkastukseen osallistuneet

[3] Itse täytettävät kyselylomakkeet palauttaneet

[4] Terveys 2000 -tutkimuksen terveyshaastatteluun, kyselyihin ja/tai terveystarkastukseen osallistuneet

Tutkimukseen poimittujen kohdehenkilöiden äidinkieli on johdettu Väestötietojärjestelmän tiedoista. Suomen ja saamen kieli on ryhmitelty samaan kieliryhmään. Ruotsin kielen äidinkielekseen rekisteröineet muodostavat yhden ryhmän. Muuta kuin suomen, saamen tai ruotsin kieltä äidinkielenään puhuvat on koottu omaksi ryhmäkseen. Heidän osuutensa koko otoksesta on pieni. Taulukosta 5.3. nähdään, että ennen painotusta muuta kieltä äidinkielenään puhuvien osuus on aliedustettu ja ruotsin kieltä puhuvien yliedustettu suhteessa väestön kielijakaumiin.

Taulukko 5.3.
Väestöjakauma kohdeperusjoukossa (30 vuotta täyttäneet) ja lopullisessa otoksessa äidinkielen mukaan

	Kohde-perusjoukko		Terveys-haastatellut		Terveys-tarkastetut		Terveys-kysely		Terveys 2000 -tutkimus	
Äidinkieli	N	%	n	%	n	%	n	%	n	%
Suomi ja saame	3 029 012	93,1	6 607	93,2	6 316	93,3	6 370	93,3	6 632	93,3
Ruotsi	172 519	5,3	405	5,7	386	5,7	391	5,7	406	5,7
Muu kieli	53 150	1,6	74	1,0	68	1,0	67	1,0	74	1,0
Kaikki	**3 254 681**	**100,0**	**7 086**	**100,0**	**6 770**	**100,0**	**6 828**	**100,0**	**7 112**	**100,0**

Asetelmapainojen kalibrointi

Alkuperäisten asetelmapainojen kalibrointi on tehty CALMAR-makrolla (Sautory, 1993). Tutkimusaineiston moniasteisesta luonteesta johtuen painokertoimia on tehty 30 vuotta täyttäneille kohdehenkilöille neljälle eri vastausjoukolle seuraavasti:

- kaikki vastanneet: osallistunut mihin tahansa tutkimuksen osioon tai erillisiin katohaastatteluihin taikka kyselyihin, n=7415
- vastanneiden unioni: osallistunut mihin tahansa tutkimuksen osioon, n=7112
- ravinto: osallistunut useimpiin osioihin ja erityisesti ravintokyselyyn, n=6005
- leikkaus: osallistunut haastatteluihin tai vastaaviin kyselyihin sekä useimpiin kliinisiin osioihin ja kyselyihin, n=5482.

On kuitenkin huomattava, että kunkin vastausjoukon sisällä voi eri tutkimusmuuttujissa tai jopa kokonaisissa osioissa syntyä erilaisia vastaajajoukkoja. Selkein esimerkki on mielenterveyttä koskeva osio, joka oli laadittu vain suomen kielellä, minkä vuoksi vastaajiksi valikoitui pelkästään riittävän hyvin kysymykset ymmärtävät henkilöt riippumatta heidän äidinkielestään. Muissa osioissa eräkatoa esiintyi pääasiassa satunnaisista syistä, muun muassa mittalaitteiden tilapäisistä häiriöistä johtuen.

Asetelmapainot laadittiin po. ryhmille erikseen. Perustana oli laajimmalle vastaajajoukolle lasketut osite- ja ryväskohtaiset sisältymistodennäköisyydet. Terveys 2000 -tutkimuksen otanta-asetelma johti melko mutkikkaaseen painojen laskentatapaan. Ensin määriteltiin rypäiden lukumääräksi m = 80 ja ne jaettiin tasakiintiöinnillä yliopistosairaalapiirien mukaisiin alueositteisiin, ts. m_h = 80/5=16. Tutkimuksen otoskoko jaettiin kuitenkin suhteellisen kiintiöinnin periaatteella, ts. $n_h = n\ (N_h / N)$. Suhteellista kiintiöintiä sovellettiin edelleen 15 suurimman kaupungin otoskokojen määrittämiseen. Niissä alkioiden sisältymistodennäköisyys on helposti palautettavissa yksinkertaisen satunnaisotannan tilanteeseen:

$\pi_{h1} = \frac{n_{h1}}{N}\frac{N}{N_1}$ ja asetelmapaino sen käänteisluku: $w_{h1} = \frac{N_1}{n_{h1}}$

Sisältymistodennäköisyydet kaksiasteisessa ryväsotannassa ovat muotoa

$$\pi_{h2} = \frac{m_{h2} n_i}{N_2}$$

missä M on rypäiden lukumäärä, N väestö, m poimittavien rypäiden lukumäärä ja n ryväskohtainen otoskoko. Alaindeksi 2 viittaa toisen asteen poimintaan, jota varten kunkin yliopistosairaalapiirin ositteista oli edellä mainittujen 15 suurimman kaupungin tiedot poistettu. Asetelmapaino on tässäkin tapauksessa sisältymistodennäköisyyden käänteisluku:

$$w_{h2} = \frac{N_2}{m_{h2} n_i}$$

Ositteiden sisällä 80-vuotiaita ja sitä vanhempia poimittiin kaksinkertaisella todennäköisyydellä muihin verrattuna. Vanhusten lukumäärän kasvattaminen ja toisen asteen rypäiden lukumäärän kiintiöintitapa kussakin yliopistosairaalapiirissä aiheutti tilanteen, jossa asetelmapainoihin tuli vaihtelua ositteiden sisällä.

Painotusvaiheessa laskettiin yllä esitetyille neljälle painotusjoukolle otospainot siten, että otoskoko korvattiin vastanneiden henkilöiden lukumäärällä ositteittain erikseen alle 80-vuotiaiden ja sitä vanhempien ryhmässä. Menettelyn tavoitteena oli pitää vastauskadosta aiheutuva painojen lisävaihtelu painotuksen alkuvaiheessa mahdollisimman harmittomana kohdistamalla sitä asetelman sisään. Kunkin painotusryhmän sisällä painokerroin oli vakio ja pääasiallinen asetelmapainojen vaihtelu aiheutui ositekohtaisista eroista alle 80-vuotiaiden ja sitä vanhempien painojen välillä.

Terveyshaastatteluaineiston painojen kalibrointiin on käytetty seuraavia otanta-asetelmasta johdettuja muuttujia ja otoshenkilöiden demografisia tietoja:

- muokattuun sisältymistodennäköisyyteen perustuva otospaino
- terveyskeskuspiiri,
- miljoonapiiri,
- ikä,
- sukupuoli ja
- äidinkieli.

Kalibroidut painokertoimet korvaavat otosaineiston painotetun jakauman edellä taulukoissa esitettyjen kohdeperusjoukon väestön jakaumien mukaiseksi.

Taulukko 5.4.
Terveys 2000 -tutkimuksen painojen keskiarvo, variaatiokerroin ja saman ryhmän painojen välinen korrelaatio

Ryhmä (n)		Keskiarvo	Variaatio-kerroin (%)	Korrelaatio
Kaikki	- asetelmapaino	438,9	15,9	
(7 415)	- kalibroitu paino	438,9	16,8	0,947
Unioni	- asetelmapaino	457,6	15,7	
(7 112)	- kalibroitu paino	457,6	16,7	0,939
Ravinto	- asetelmapaino	542,0	9,5	
(6 005)	- kalibroitu paino	542,0	13,8	0,692
Leikkaus	- asetelmapaino	593,6	9,6	
(5 482)	- kalibroitu paino	593,6	17,6	0,325

Kalibroinnissa asetelmaan on tuotu lisää informaatiota, minkä vuoksi painojen vaihtelu lisääntyi. Kahdessa ensimmäisessä joukossa vastaajien ja otoksen väliset jakaumat ovat suhteellisen lähellä toisiaan, joten muutos ei ollut kovin suuri. Ravintotutkimukseen osallistuneiden sekä leikkausjoukkoon kuuluneiden osalta lisätiedolla oli sen sijaan suurehko vaikutus. Näissä ryhmissä vanhusten osuus vastanneista pieneni jyrkästi, mikä näkyy asetelmapainojen variaatiokertoimien pienuutena. Kun kalibroinnissa ryhmien suhteet palautetaan oikeiksi eri taustamuuttujien suhteen, painojen vaihtelu lisääntyy merkittävästi.

Koska kalibroinnissa käytettävät muuttujat voivat vaikuttaa painotuksen kautta lopullisiin estimaatteihin, niiden valinnassa oltiin erityisen kriittisiä Terveys 2000 -tutkimuksen tapauksessa. Aineistoa tullaan hyödyntämään useisiin eri tutkimuksiin ja niissäkin hyvin monimuotoisiin ja monimutkaisiin analyysiasetelmiin. Jotta aineiston käyttökelpoisuus ja vertailukelpoisuus säilyy hyvänä yli ajan, on kalibroinnissa pyritty käyttämään vain selkeitä demografisia muuttujia, eikä johdettuja (ja arvottavia) muuttujia, kuten esimerkiksi sosioekonominen asema. Painokertoimien kalibroinnissa tulee myös välttää liian suurta painotussolujen lukumäärää, minkä vuoksi esimerkiksi kieliryhmiä yhdistettiin ja ikäluokat pidettiin riittävän suurina.

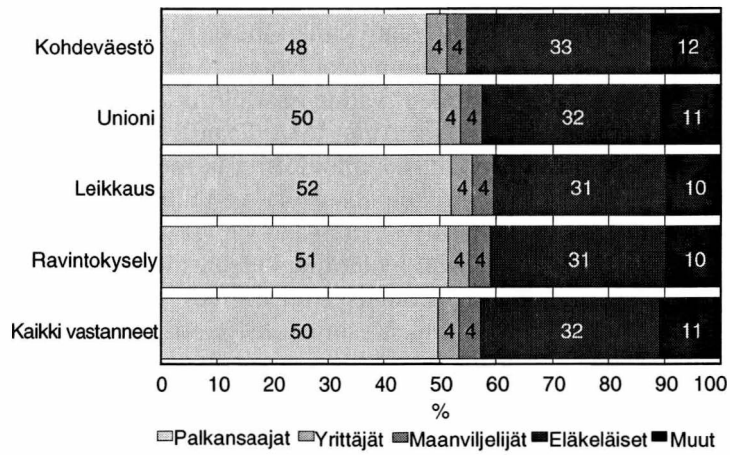
Seuraavassa verrataan painotettuja otosjakaumia kohdeväestöön (Laiho, 2002b). Kuvioista 5.1.–5.3. huomataan, että kalibroidut painokertoimet yleistävät otosväestön erittäin hyvin kohdeväestön tasolle myös karkean sosioekonomisen aseman, siviilisäädyn ja suuralueen mukaan, vaikka näitä muuttujia ei ole käytetty painokertoimien kalibroinnissa. Kyseiset luokitukset on selitetty tarkemmin liitteessä: Laatuselvityksessä käytetyt käsitteet ja määritelmät.

Tarkasteltaessa sosioekonomisen aseman jakaumaa hyväksytyssä lopullisessa Terveys 2000 -aineistossa (eli otosväestössä) ja kohdeväestössä havaitaan, että palkansaajien ja yrittäjien suhteellinen osuus on otosväestössä hieman suurempi kuin kohdeväestössä. Sitä vastoin eläkeläisten, opiskelijoiden ja muiden

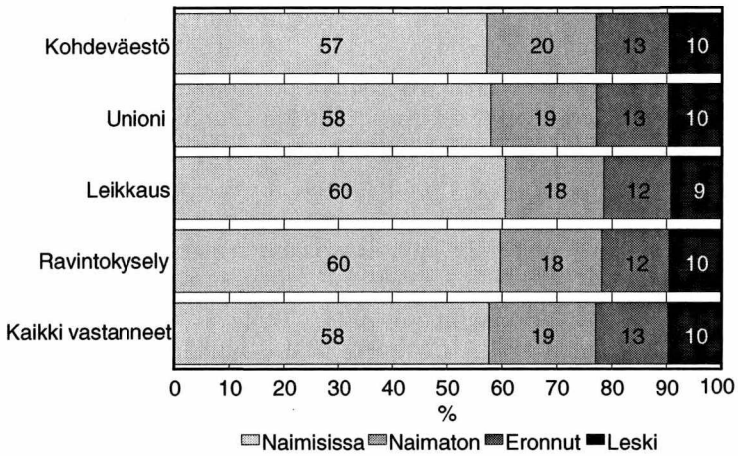
osuus on otosväestössä hieman alhaisempi kuin kohdeväestössä. Siviilisäädyn mukaisissa tarkasteluissa merkittävimmät erot otos- ja kohdeväestön välillä on avioituneiden osuudessa, joiden osuus on otosväestössä suhteellisesti hieman korkeampi kuin kohdeväestössä. Vastaavasti naimattomien, eronneiden ja leskien osuus on suhteellista alhaisempi otosväestössä kuin kohdeväestössä. Nämä erot kohdeväestön ja painotettujen otosjakaumien välillä eivät kuitenkaan ole tilastollisesti merkitseviä sosioekonomisen aseman taikka siviilisäädyn mukaan. Otosjakaumien painotus korjaa hyvin otosjakaumat myös niissä väestöryhmissä, joissa yksinasuvia on suhteellisesti paljon (naimattomat, eronneet ja lesket). Kuten edellä on todettu, yksinasuvien kato-osuudet olivat erityisen suuria.

Painokertoimet on kalibroitu siten, että miljoonapiirien mukainen väestöjakauma on painotetussa otosväestössä sama kuin kohdeväestössä. Kuten edellä on todettu, miljoonapiirit ovat terveystutkimukselle tarkoituksenmukaisempi alueellinen luokitus kuin NUTS-luokitukseen perustuvat suuralueet. Tällöin suuralueiden mukaisessa tarkastelussa on havaittavissa pieniä eroja, etenkin jos pääkaupunkiseutu on eriytetty muusta Etelä-Suomesta. Vaikka erot kohdeväestön ja otosväestöjen välillä ovat marginaalisia, on Etelä-Suomessa asuvien osuus hieman otosväestössä aliedustettu ja Pohjois-Suomen väestön osuus vähän yliedustettu.

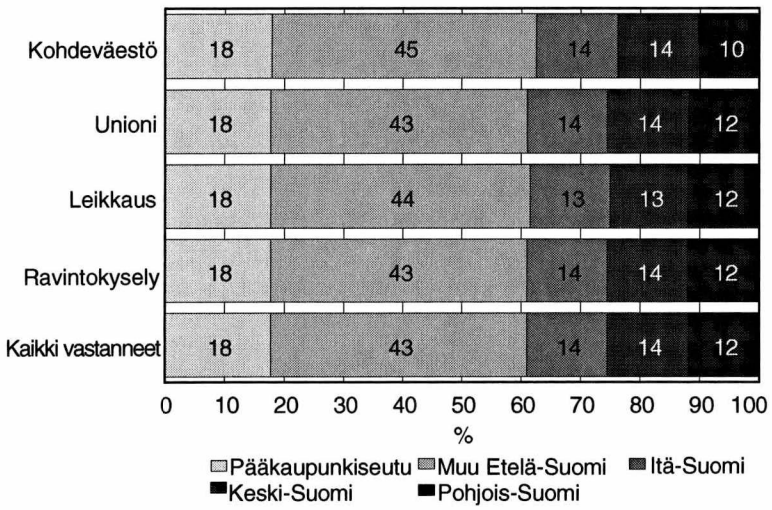
Kuvio 5.1.
Väestöjakauma kohdeperusjoukossa (30 vuotta täyttäneet) ja lopullisessa otoksessa (painotettu) sosioekonomisen aseman mukaan



Kuvio 5.2.
Väestöjakauma kohdeperusjoukossa (30 vuotta täyttäneet) ja lopullisessa otoksessa (painotettu) siviilisäädyn mukaan



Kuvio 5.3.
Väestöjakauma kohdeperusjoukossa (30 vuotta täyttäneet) ja lopullisessa otoksessa (painotettu) suuralueen mukaan



Suosituksia painojen käyttämisestä

Edellä esitetyt neljä painokerrointa korottavat otoshavainnot perusjoukon tasolle. Niiden ohella tarvitaan painokertoimet, joiden tilastolliset ominaisuudet pysyvät ennallaan, mutta joiden summa on vastanneiden lukumäärä ja keskiarvo on 1. Jokaiselle painokertoimista on johdettu tällainen analyysipainoksi kutsuttu paino. Useimmissa keskiarvoihin, osuuksiin tai muihin sellaisiin tunnuslukuihin ja erilaisiin mallituksiin perustuvissa tarkasteluissa on suositeltavaa käyttää juuri analyysipainoa ja vastaavasti kokonaismäärien estimoinnissa väestötason painoa.

On vaikea antaa suositusta, mikä johdetuista painokertoimista olisi paras, koska Terveys 2000 -tutkimuksen aineistoa käytetään erittäin monenlaisiin tarkoituksiin. Nyrkkisääntönä olkoon, että tutkittavan joukon ominaisuuksien pitäisi olla mahdollisimman lähellä jotain edellä mainituista joukoista, jotta painotus olisi mahdollisimman oikea. Esimerkiksi ravintopäiväkirjaa koskevissa tutkimuksissa tulisi aina käyttää tälle joukolle laskettua painoa. Ja vastaavasti useiden muuttujien yhdistelmää tarkasteltaessa eri muuttujissa oleva eräkato kumuloituu, jolloin tutkimusjoukko saattaa olla lähinnä leikkauspainojen laskennassa mukana ollut joukko. Maksimaalista osallistumista kuvaavaa painoa "kaikki" ei kannattane käyttää kuin rekistereistä johdettua ja lähinnä taustatietoja koskevassa analyysissä. Siten useimmissa tapauksissa lähinnä oikein painokerroin on joko unionipaino tai leikkauspaino.

5.2 Otosvarianssin ja keskivirheen estimointi

Terveys 2000 -tutkimuksen otannassa ja estimoinnissa sovellettua asetelmaa voidaan kutsua monimutkaiseksi, koska suuri osa alkioista poimittiin käyttäen kaksiasteista ositettua otantaa. Monimutkaisten otanta-asetelmien tapauksessa piste-estimaattien otosvarianssin laskentakaavojen johtaminen edellyttää otosasetelman kunnollista huomioon ottamista. Edellä kappaleessa 3.2 on kuvattu tutkimuksen otanta-asetelmaa. Varianssiestimaattoreiden johtamiseksi aineistoa on muokattu siten, että 15 suurimmassa kaupungissa sovellettu yksiasteinen otanta erotetaan kaksiasteisesta otannasta oikeiden varianssiestimaattoreiden johtamiseksi. Seuraavassa menettelyä on kuvattu tarkemmin.

15 suurimman kaupungin otanta-asetelma oli yksiasteinen ositettu alkiotason otanta. Sen vuoksi tutkimusaineistoihin on tehty tekninen muokkaus, jolla jokainen näistä kaupungeista muodostaa oman ositteensa ja edelleen jokaista tutkimukseen vastannutta henkilöä käsitellään rypäänä. Esimerkiksi terveyshaastatteluun osallistuneita oli näissä kaupungeissa yhteensä 2 695, jolloin varianssiestimaattorissa vapausasteiden lukumääräksi muodostuu koko maassa 2 680 (2 695 ryvästä[1]/alkiota -15 ositetta).

Varsinaisessa kaksiasteisessa osassa on kullakin ositteella oma miljoonapiiritunnuksensa: ensimmäinen otanta-aste muodostuu terveyskeskuspiirin ja toinen aste henkilöiden poiminnasta. Rypäille on annettu juokseva numerointi,

[1] Otokseen poimittuja henkilöitä, joita tutkimusasetelmasta johtuen käsitellään rypäinä.

jotta käsittely tulisi mahdollisimman yksinkertaiseksi. Näitä PPS-tyyppisellä otannalla poimittuja rypäitä oli koko maassa 65, jolloin varianssiestimaattorin kaavassa vapausasteiden lukumääräksi muodostuu rypäiden lkm - ositteiden lkm eli 65-5 = 60.

Asetelmapohjaista estimointia kuvataan moderneissa otantateorian oppikirjoissa, erityisesti Lehtosen ja Pahkisen (1996), Skinnerin, Holtin ja Smithin (1989) ja Lohrin (1999) teoksissa, samoin erään yleisimmin käytetyn analyysiohjelman käsikirjassa (Research Triangle Institute, 2001).

Monimutkaisille otanta-asetelmille otosvarianssin analyyttinen johtaminen ei ole aina mahdollista. Myös estimoitavat parametrit tai niiden yhdistelmät esimerkiksi osajoukkojen tasolla ovat usein epälineaarisia. Moniasteisen ryväsotannan tapauksessa joudutaan käyttämään approksimaatiota, joka voi perustua joko Taylorin sarjakehitelmään tai otoksen uudiskäyttötekniikoihin. Menetelmiin viitataan kaikissa edellä mainituissa teoksissa.

Seuraavassa esitetään useimmiten käytetty approksimaatio, jossa oletetaan, että otos on poimittu palauttaen, mutta vaihtelevin todennäköisyyksin. Approksimaatio ottaa huomioon vain rypäiden välisen varianssin, mutta jättää rypäiden sisäisen varianssin estimoinnin ulkopuolella. Menetelmä edellyttää, että kustakin ositteesta on poimittu vähintään kaksi ryvästä ja että poimintasuhde eli poimittujen rypäiden osuus ositteen kaikista rypäistä on pieni. Asetelmasta tarvitaan seuraavat tiedot: osite, ryväs ja paino (joko korottava tai analyysipaino), estimoitavaksi parametriksi oletetaan minkä tahansa tutkimusmuuttujan keskiarvo.

Keskiarvon estimaattori koko perusjoukolle on

$$\hat{\bar{Y}} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} y_{hij} \Big/ \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij}$$

Taylorin sarjakehitelmään perustuva varianssiestimaattori keskiarvolle voidaan esittää muodossa:

$$\hat{V}(\hat{\bar{Y}}) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \frac{m_h}{m_h - 1} \left(Z_{hi} - \bar{Z}_h\right)^2$$

missä $Z_{hi} = \sum_{j=1}^{m_{hi}} w_{hij} \left(y_{hij} - \hat{\bar{Y}}\right) \Big/ \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij}$,

ja $\bar{Z}_h = \sum_{i=1}^{n_h} Z_{hi} / m_h$.

Keskivirheen estimaattori on varianssiestimaattorin neliöjuuri, ts.
$StdErr(\hat{Y}) = \sqrt{\hat{V}(\hat{Y})}$.

5.3 Tehokkuusvertailu

Otantatutkimuksissa käytetyn asetelman tehokkuutta vertaillaan usein yksinkertaista satunnaisotantaa käyttävää asetelmaa vastaan. Vertailu tehdään suhdeluvulla, jota kutsutaan asetelmakertoimeksi:

$$deff(\hat{\theta}) = \frac{\hat{V}(\hat{\theta})_{p(s)}}{\hat{V}(\hat{\theta})_{srswr}}$$

Ryväsotantaa käyttävissä asetelmissa on tyypillistä, että havainnot ovat rypäiden sisällä sisäkorreloituneita. Toisin sanoen rypäiden sisällä olevien henkilöiden välinen vaihtelu on vähäisempää kuin se olisi vastaavan kokoisessa otoksessa suoraan perusjoukosta poimittuna. Mikäli asetelmakerroin saa arvon 1, käytetty otanta-asetelma on yhtä tehokas kuin yksinkertainen satunnaisotanta. Ykköstä pienempi arvo osoittaa, että asetelma on tehokkaampi kuin yksinkertainen satunnaisotanta. Ryväsotannan tapauksessa asetelmakertoimet ovat usein ykköstä suurempia, mikäli rypäisiin valikoitumisen ja tutkimuksen kohteena olevan ilmiön välillä on riippuvuutta. Joskus sellaista riippuvuutta ei ole, vaan tutkittavan muuttujan vaihtelu "lävistää" otosasetelman.

Seuraavassa taulukossa 5.5. on esitetty muutamia Terveys 2000 -tutkimuksen muuttujia taulukoituna otosasetelman monimutkaisuuden suhteen. Oletuksen mukaisesti yksiasteisesta asetelmasta estimoidut asetelmakertoimet ovat lähellä ykköstä, sen sijaan kaksiasteisesta asetelmasta estimoidut asetelmakertoimet vaihtelevat muuttujasta toiseen. Taulukko osoittaa, ettei tutkittavien muuttujien riippuvuutta otosasetelmasta voida sulkea pois. Sen vuoksi on turvallisinta, että aineistosta tehtävissä analyyseissä käytetään sellaisia ohjelmistoja, jotka kykenevät ottamaan tutkimuksen otosasetelman huomioon, esimerkiksi SUDAAN tai STATA taikka kappaleen 6 kaltaisilla malleilla, joissa aineiston hierarkkisuus voidaan ottaa huomioon mallia rakennettaessa.

Taulukko 5.5.
Eräiden Terveys 2000 -tutkimuksen muuttujien asetelmakertoimet otanta-asetelman suhteen. Otospainona on käytetty vastaajien lukumäärään skaalattua unionipainoa.

Tutkimus-muuttuja	Otos-asetelma	Havaintojen lkm	Keski-arvo	Keski-virhe	Asetelma-kerroin
Diastolinen verenpaine	kaikki	6 334	81,5	0,29	4,4
mmHg	1-asteinen	2 468	81,2	0,21	1,0
	2-asteinen	3 866	81,7	0,46	6,4
Systolinen verenpaine	kaikki	6 336	133,5	0,43	2,7
mmHg	1-asteinen	2 468	131,1	0,40	1,0
	2-asteinen	3 868	135,0	0,65	3,7
Lääkärissäkäynnit	kaikki	6 986	3,1	0,06	1,2
sairauden vuoksi	1-asteinen	2 695	3,3	0,08	1,0
	2-asteinen	4 291	2,9	0,08	1,3
Itsearvioitu	kaikki	6 981	61,0	0,63	1,2
terveydentila	1-asteinen	2 693	65,0	0,92	1,0
hyvä, %	2-asteinen	4 288	58,5	0,85	1,3
Pitkäaikainen	kaikki	6 981	52,8	0,77	1,7
sairaus	1-asteinen	2 693	47,9	0,96	1,0
%	2-asteinen	4 288	55,9	1,10	2,1
Painoindeksi	kaikki	5 979	26,6	0,06	1,2
(BMI)	1-asteinen	2 445	26,1	0,09	1,0
	2-asteinen	3 534	26,9	0,08	1,2
Vyötärön	kaikki	6 289	92,9	0,18	1,2
ympärys	1-asteinen	2 449	91,5	0,27	1,0
cm	2-asteinen	3 840	93,8	0,25	1,4

5.4 Mallivakiointiin perustuva estimointi

Sukupuolen, alueen tai muun luokitellun muuttujan perusteella määriteltyjen osajoukkojen välisten erojen arvioimista saattaa vaikeuttaa muiden tekijöiden, kuten iän, erilainen jakautuminen em. osajoukoissa. Erot esimerkiksi keskiarvoissa tai prevalensseissa voivat johtua yksinomaan muiden tekijöiden vaikutuksesta, tai muut tekijät saattavat peittää todellisen eron. Muiden tekijöiden vaikutukset voidaan vakioida käyttämällä Leen (1981) menetelmällä ennustettuja reunajakaumia (*predictive margins*). Tässä menetelmässä estimoidaan regressiomalli, jossa vasteena on tutkittava muuttuja ja selittäjänä osajoukkomuuttujan lisäksi muut tekijät, joiden oletetaan vaikuttavan tutkittavan muuttujan arvoon.

Jokaiselle henkilölle lasketaan mallin perusteella yksilöllinen ennuste asettamalla osajoukkomuuttujan arvo samaksi kaikille henkilöille. Muiden se-

littäjien arvot pidetään ennallaan. Lopputuloksena raportoidaan yksilöllisten ennusteiden keskiarvot. Graubard ja Korn (1999) ovat johtaneet mallivakioiduille tunnusluvuille keskivirhe-estimaattorit.

Esimerkkinä on malli, jossa tulosmuuttujana on systolinen verenpaine ja selittäjinä painoindeksi BMI, sukupuoli, karkea ikäryhmitys, kokonaiskolesteroli ja siviilisääty. Esimerkkiaineisto on ollut käytettävissä vain Kansanterveyslaitoksessa, joten taulukoissa esiintyvät totaalit poikkeavat tässä julkaisussa muualla esitetyistä. Mallivakiointi on tehty taulukossa 5.6. sukupuolen mukaan, eli olettaen kaikkien olevan joko naisia tai miehiä. Erot ovat melko pienet sukupuolten välillä myös havaituissa keskiarvoissa, mutta jos mallissa sukupuolen ja iän välille asetetaan yhdysvaikutus, niin mallivakiointi taulukossa 5.7. – olettaen kaikkien olevan joko naisia tai miehiä ja kuuluvan samaan ikäluokkaan – antaa selvimmän eron nuorten naisten ryhmässä verrattaessa havaittuun keskiarvoon.

Taulukko 5.6.
Mallivakiointiesimerkin tulokset sukupuolittain. Vasteena systolinen verenpaine.[1]

Sukupuoli	Ennustettu reunajakauma	Keskivirhe	Solufrekvenssi	Havaittu keskiarvo	Keskihajonta
1	134,5	0,4	3 008	134,6	19,2
2	132,0	0,5	3 639	132,8	22,8

Taulukko 5.7.
Mallivakiointiesimerkin tulokset sukupuolittain ja ikäryhmittäin. Vasteena systolinen verenpaine.[1]

Sukupuoli	ikä (6 lk.)	Ennustettu reunajakauma	Keskivirhe	Solufrekvenssi	Havaittu keskiarvo	Keskihajonta
1	1	124,3	0,6	706	124,8	13,1
1	2	129,6	0,7	755	130,1	15,6
1	3	135,0	0,7	696	136,8	18,9
1	4	143,3	1,0	468	144,5	21,2
1	5	145,1	1,5	254	145,9	21,2
1	6	145,3	2,6	129	144,8	21,4
2	1	117,7	0,6	793	115,7	12,8
2	2	124,4	0,7	847	124,2	16,9
2	3	134,1	0,8	751	135,2	20,0

[1] Huom. Taulukoiden 5.6 ja 5.7 otoskoot eivät ole samoja kuin muissa tämän laaturaportin taulukoissa ja analyyseissä. Taulukoiden 5.6 ja 5.7 analyysit perustuvat KTL:n lopulliseen tarkistettuun aineistoon, joka ei ole ollut Tilastokeskuksessa käytettävissä muita analyysejä suoritettaessa.

6 Monimuuttuja-analyysi: vertailuja ja suosituksia

Risto Lehtonen, Kari Djerf, Tommi Härkänen ja Johanna Laiho

Tässä jaksossa havainnollistetaan Terveys 2000 -tutkimuksen aineiston käyttöä monimuuttujaisissa analyysitilanteissa. Esimerkkianalyysien eri menetelmillä otetaan huomioon otanta-asetelman ominaisuuksia ja verrataan tuloksia analyysiin, jossa otanta-asetelmaan ei reagoida. Analyyseissa käytetään menetelmiä, joiden avulla analyysiproseduuriin tuodaan mukaan painotus (painomuuttuja) sekä asetelman ositus (ositeindikaattori) ja ryvästyminen (ryväsindikaattori).

Painomuuttujan avulla kompensoidaan sisältymistodennäköisyyksien vaihtelu (80 vuotta täyttäneiden yliedustus) sekä oikaistaan vastauskadon mahdollisesti tuottamaa harhaa. Painomuuttujan käytön tarkoitus on tuottaa tilastollisessa mielessä tarkentuvia piste-estimaatteja (esimerkiksi keskiarvot, esiintyvyydet ja regressiokertoimien estimaatit). Otanta-asetelmaan sisältyvä alueellinen ryvästys tuotti useille tulosmuuttujille positiivista sisäkorreloituneisuutta rypäissä. Tällöin tulosmuuttujien piste-estimaattien asetelmakertoimet ovat ykköstä suurempia, eli asetelmaperusteisesti estimoidut keskivirheet ovat suurempia kuin keskivirheet, jotka lasketaan olettamalla aineiston perustuvan yksinkertaiseen satunnaisotantaan suoraan henkilöperusjoukosta.

Painotuksen, osituksen ja ryvästymisen huomioon ottamista varten analyyseissa on vaihtoehtoisia lähestymistapoja ja menetelmiä. Menetelmät voidaan jakaa karkeasti asetelmaperusteisiin menetelmiin ja malliperusteisiin menetelmiin. Esityksessä keskitytään asetelma- ja malliperusteisiin menetelmiin, jotka mahdollistavat painotuksen ja ryvästymisen huomioon ottamisen analyyseissa. Otanta-asetelmaan sisältyvä alueellinen ositus tuodaan mukaan analyysiproseduuriin eräiden asetelmaperusteisten analyysien yhteydessä.

Menetelmiin liittyvät tilastolliset mallit käsittävät yleistettyjen lineaaristen mallien ja lineaaristen sekamallien sovelluksia. Käytettäviä estimointimenetelmiä ovat painotettu pienimmän neliösumman (PNS) estimointi, pseudouskottavuusmenetelmä (PML, *pseudo maximum likelihood*) ja yleistettyjen estimointiyhtälöiden menetelmä (GEE, *generalized estimating equations*) sekä lineaaristen sekamallien yhteydessä yleistetty PNS-menetelmä (GLS; *generalized least squares*) ja REML-menetelmä (*residual maximum likelihood*).

SUDAAN-ohjelmisto on tärkein asetelmaperusteisiin menetelmiin erikoistunut ohjelmatuote. Myös SAS-ohjelmistossa on työkaluja, joilla voidaan päätyä likimain samoihin numeerisiin tuloksiin kuin SUDAAN-ohjelmistolla. Molemmissa tapauksissa käyttäjän tulee osata tehdä oikeita valintoja ohjelmistojen asianomaisten proseduurien analyysiasetelmien valikoimasta. Tässä jaksossa esitellään ohjelmistojen proseduurien käyttöä lineaarisen ja logistisen kovarianssianalyysin yhteydessä. Esitys perustuu artikkeleihin Lehtonen, Djerf, Härkänen ja Laiho (2003a ja b).

6.1 Otanta-asetelma ja asetelmakertoimet

Aineistossa on 20 ositetta, joista 15 edustaa Manner-Suomen suurimpia kaupunkeja. Näiden kaupunkien terveyskeskuspiirit on poimittu otokseen todennäköisyydellä yksi. Muut 5 ositetta kattavat jäljelle jäävän osan maan viidestä miljoonapiiristä. Nämä viisi alueellista ositetta voidaan jakaa edelleen 234 terveyskeskuspiiriin, joista otokseen on poimittu kaikkiaan 65. Kaikkiaan otosrypäitä on 80.

Aineiston koko tässä jaksossa esiteltävissä analyyseissa on 5 954 henkilöä. Ensimmäisen asteen poimintayksikköjä on 2 495, joista 65 on kaksiasteisista ositteista ja 2 430 yksiasteisista ositteista. Siten asetelmaperusteisia analyyseja varten on käytettävissä kaikkiaan 2 475 vapausastetta (*denominator degrees of freedom*). Analyysipaino on konstruoitu edellisessä luvussa kuvatulla tavalla. Tämän luvun esimerkeissä analyysipainot on skaalattu niin, että niiden keskiarvo yli aineiston on 1,0. Painojen vaihtelua kuvaava variaatiokerroin on noin 15 %, mikä on varsin pieni.

Analyysia varten valittiin tulosmuuttujiksi systolinen verenpaine ja pitkäaikaissairastavuus. Selittäväksi muuttujaksi valittiin vyötärönympärys (CIRCUM) systoliselle verenpaineelle ja koulutus (vuosina) pitkäaikaissairastavuusmuuttujalle. Molemmat mallit vakioitiin sukupuolen ja iän suhteen.

Tulosmuuttujien asetelmakertoimet (*design effect*, deff) on esitetty edellä taulukossa 5.5. Asetelmakertoimet olivat selvästi ykköstä suurempia. Painotuksen vaikutus asetelmakertoimiin oli vähäinen. (Lehtonen et al. 2003a ja b).

6.2 Analyysimenetelmät

Yleistettyjen lineaaristen mallien sovittamisessa käytettiin ns. häiriötekijälähestymistapaa (*nuisance approach*; Skinner et al. 1989; Lehtonen ja Pahkinen 1996), joka tuottaa asetelman suhteen tarkentuvan estimoinnin ja asymptoottisesti pätevän tilastollisen testauksen. Estimoinnissa ja testauksessa käytettiin pseudo-uskottavuusmenetelmää ja asetelmaperusteisia keskivirhe-estimaattoreita sekä asetelmaperusteisia Waldin testisuureita.

Yleistettyjen estimointiyhtälöiden GEE-menetelmää (Liang ja Zeger 1986; Diggle, Liang ja Zeger 1994) käytettiin toisena asetelmaperusteisena vaihtoehtona. GEE-menetelmästä käytettiin tyypiltään vaihdettavan *(exchangeable)* sisäkorrelaatiorakenteen versiota. Siinä sallitaan, että rypään henkilöiden pareittainen sisäkorrelaatio voi poiketa nollasta, mutta oletetaan vakioksi kaikissa rypäissä.

Kolmas menetelmävaihtoehto perustui ns. sekamalleihin *(mixed models, multilevel models;* Goldstein 1995; McCulloch and Searle 2001), joissa on kiinteiden tekijöiden lisäksi ryvästasoisia satunnaistekijöitä. Menetelmien empiirisessä vertailussa laskettiin mallien parametrien piste-estimaatit, asetelmaperusteiset keskivirheet, asetelmakertoimet sekä asetelmaperusteisia t-testisuureita eli Waldin testisuureen F-korjattuja merkkisiä neliöjuuria.

Menetelmien vertailua varten määriteltiin joukko analyysiasetelmia (taulukko 6.1). Asetelmassa 1 käytettiin PML-menetelmää yleistettyjen lineaaristen mallien parametrien estimointiin. Menetelmä vastaa GEE-menetelmää, jossa oletetaan riippumaton korrelaatiorakenne rypäiden sisällä. PML-menetelmä palautuu painotettuun pienimmän neliösumman (*weighted least squares,* WLS) estimointiin lineaaristen mallien asetelmaperusteisessa estimoinnissa, ja tavanomaiseen painotettuun pienimmän neliösumman menetelmään, jos painokertoimet oletetaan ykkösiksi.

Asetelmassa 2 käytettiin GEE-menetelmän vaihdettavan sisäkorrelaation versiota. Asetelman 1 mukaisissa PML- ja WLS-menetelmissä otanta-asetelman ryvästyminen vaikuttaa vain mallin kerroinestimaattoreiden keskivirheiden estimointiin, kun taas asetelman 2 mukaisessa GEE-analyysissa ryvästyminen vaikuttaa myös piste-estimaattien laskentaan. Molemmissa analyysiasetelmissa keskivirhe-estimaatit laskettiin asetelmaperusteisella *sandwich*-estimaattorilla (Lehtonen ja Pahkinen 1996, s. 271). Vertailun vuoksi asetelman 2 mukainen analyysi tehtiin sekä SUDAAN-ohjelmiston proseduureilla REGRESS ja LOGISTIC/RLOGIST (asetelma 2a) että SAS-ohjelmiston GENMOD-proseduurilla (asetelma 2b).

Asetelma 3 perustuu lineaariseen sekamalliin, joka määriteltiin varianssikomponenttimallina. Siihen sisältyy mallin kiinteiden tekijöiden lisäksi ryvästasoiset satunnaiskomponentit (*random intercepts*). Varianssikomponentit estimoitiin REML-menetelmällä ja kiinteiden tekijöiden kertoimet yleistetyllä PNS-menetelmällä (*generalized least squares*, GLS) ehdolla satunnaistekijöiden estimaatit. Mallit sovitettiin SAS-ohjelmiston MIXED-proseduurilla. Asetelmissa 1, 2 ja 3 analyysipainot tuotiin mukaan analyyseihin painomuuttujan avulla. Painotusta käytettiin asetelman suhteen tarkentuvan estimoinnin saavuttamiseksi sekä suojaamaan mahdollisen virheellisen mallin määrityksen tuottamalta harhaisuudelta (esim. Pfeffermann et al. 1998).

Analyysiasetelma 0 perustuu oletukseen, että aineisto on saatu yksinkertaisella satunnaisotannalla, eikä analyysipainoja käytetä estimoinnissa. Tämä oletus sivuuttaa kaikki käytetyn otanta-asetelman ominaisuudet. Yleistettyjen lineaaristen mallien parametrit estimoidaan tällöin ML-menetelmällä (logitmalli) tai PNS-menetelmällä (lineaarinen malli).

Taulukko 6.1.
Analyysiasetelmat

		Otanta-asetelman ominaisuudet, joihin pyritään reagoimaan		
Asetelma	**Malli ja estimointimenetelmä**	**Painotus**	**Ositus**	**Ryvästys**
0	Vertailuasetelma. Kiinteiden tekijöiden malli; ML- tai LS-estimointi, malliperusteiset keskivirheet	Ei	Ei	Ei
1	Kiinteiden tekijöiden malli; PML- tai WLS-estimointi, asetelmaperusteiset keskivirheet	Kyllä	Kyllä	Kyllä
2 a) ja b)	Kiinteiden tekijöiden malli; QML-estimointi, asetelmaperusteiset keskivirheet a) SUDAAN, b) SAS/GENMOD	Kyllä	a) Kyllä b) Ei	Kyllä
3	Sekamalli; REML- ja GLS-estimointi, asetelmaperusteiset keskivirheet	Kyllä	Ei	Kyllä

Lyhenteet:
ML: Maximum likelihood
PML: Pseudo maximum likelihood
QML: Quasi maximum likelihood
REML: Residual maximum likelihood

LS: Least squares
WLS: Weighted least squares
GLS: Generalized least squares

Taulukkoon 6.2. on koottu keskeisimpien käytettävissä olevien tilastollisten ohjelmatuotteiden ominaisuuksia. Horton ja Lipsitz (1999) ovat vertailleet GEE-menetelmien käyttöä SAS, SUDAAN, STATA ja S-Plus-ohjelmistoilla. GEE-menetelmiä on lisäksi selvitetty artikkelissa Ziegler et al. (1998) ja sekamallien sovittamista MIXED-proseduurilla artikkelissa Singer (1998).

Taulukko 6.2.
Ohjelmistot ja niiden ominaisuuksia

Ohjelmisto	Asetelmavalinnat	Tulosmuuttujan tyyppi	Analyysi-proseduuri
SUDAAN (versio 8.0.1)	**Asetelma 1** GEE, "independent" korrelaatiorakenne. Vastaa PML-estimointia	Jatkuva	REGRESS
		Binäärinen	RLOGIST
		Moniluokkainen	MULTILOG
		Lukumäärämuuttuja	LOGLINK
	Asetelma 2a GEE, vaihdettava korrelaatiorakenne	Jatkuva	REGRESS
		Binäärinen	RLOGIST
		Moniluokkainen	MULTILOG
		Lukumäärämuuttuja	LOGLINK
SAS (versio 8.2)	**Asetelma 1** GEE, riippumaton korrelaatiorakenne. Vastaa PML-estimointia	Jatkuva	GENMOD
		Binäärinen	GENMOD
		Moniluokkainen	GENMOD
		Lukumäärämuuttuja	GENMOD
	Asetelma 2b GEE, vaihdettava korrelaatiorakenne	Jatkuva	GENMOD
		Binäärinen	GENMOD
		Moniluokkainen	(ei ole)
		Lukumäärämuuttuja	GENMOD
	Asetelma 3 REML, sekamallit	Jatkuva	MIXED
		Binäärinen	NLMIXED
		Moniluokkainen	GLIMMIX
		Lukumäärämuuttuja	GLIMMIX

6.3 Analyysimenetelmien empiirinen vertailu

Systolinen verenpaine

Systolisen verenpaineen vaihtelun tarkastelussa päämielenkiinto oli vyötärön ympärysmittaa kuvaavan muuttujan selitysvoimassa. Muuttujaa käytettiin havainnollisuuden vuoksi kolmiluokkaiseksi luokiteltuna muuttujana RCIRCUM. Luokittelu on tehty niin, että kussakin kolmannesluokassa on likimain sama lukumäärä henkilöitä. Analyysissa vakioitiin sukupuolen ja iän vaikutus.

Analyysiasetelma on kuvattu taulukossa 6.3. Havainnollisuuden vuoksi ikä on esitetty taulukossa kolmiluokkaisena muuttujana, mutta varsinaisessa analyysissä ikää käytettiin jatkuvana muuttujana. Taulukon mukaan systolisen verenpaineen taso vaihteli selvästi selittävien muuttujien muodostaman luokituksen mukaan. Verenpaineen taso nousi ikävuosien lisääntyessä ja kussakin ikäluokassa taso nousi myös vyötärönympäryksen kasvaessa. Osajoukkojen keskiarvojen asetelmakertoimien estimaatit olivat yleensä ykköstä suurempia, mikä viittaa lievään positiiviseen sisäkorreloituneisuuteen rypäissä.

Taulukko 6.3.
Keskimääräinen systolinen verenpaine sukupuolen, iän ja vyötärönympäryksen (RCIRCUM) mukaan

Miehet ikäryhmittäin muuttujan RCIRCUM mukaan				
	n	Keskiarvo	Keskivirhe	.Asetelma-kerroin
30–45 vuotta				
1 (alin)	232	120,7	0,9	1,1
2	450	124,4	0,7	1,5
3 (ylin)	342	131,0	0,8	1,2
46–64 vuotta				
1 (alin)	119	129,3	1,8	1,1
2	426	135,1	0,9	1,0
3 (ylin)	666	140,2	0,7	1,0
65 vuotta				
1 (alin)	47	135,9	2,9	0,9
2	170	145,9	2,3	1,7
3 (ylin)	258	146,6	1,4	1,2
Kaikki	2 710	134,4	0,5	1,9
Naiset ikäryhmittäin muuttujan RCIRCUM mukaan				
	n	Keskiarvo	Keskivirhe	Asetelma-kerroin
30–45 vuotta				
1 (alin)	782	115,6	0,5	1,1
2	222	119,6	0,9	1,0
3 (ylin)	147	127,8	1,5	1,2
46–64 vuotta				
1 (alin)	567	129,0	0,8	1,0
2	393	135,5	1,1	1,1
3 (ylin)	339	142,1	1,2	1,3
65 vuotta				
1 (alin)	228	148,6	1,5	1,1
2	282	148,8	1,4	1,1
3 (ylin)	261	152,2	1,3	0,9
Kaikki	3 220	132,2	0,5	1,8

Systolisen verenpaineen vaihtelua mallinnettiin lineaarisen ANCOVA-mallin avulla, jossa termeinä olivat kaikkien selittäjien päävaikutukset ja pareittaiset yhdysvaikutukset. Eri analyysiasetelmien antamat tulokset on koottu liitetaulukkoon 1. Kaikkien analyysiasetelmien mukaan sopiva malli sisältäisi kaikki päävaikutukset ja vyötärönympäryksen ja iän yhdysvaikutuksen. Kuitenkin asetelman 0 mukainen analyysi (jossa oletetaan yksinkertainen satunnaisotanta ja painoja ei käytetä) antaisi enemmän selitysvoimaa yhdysvaikutustermille kuin analyysiasetelmat, joissa puhdistetaan rypäiden positiivisen sisäkorreloituneisuuden vaikutus (taulukko 6.4.).

Taulukko 6.4.
Yhdysvaikutustermin RCIRCUM*AGE asetelmaperusteiset testitulokset

	DF	F-testi	p-arvo
Asetelma 0	2	5,6	0,004
Asetelma 1	2	5,2	0,005
Asetelma 2 a	2	4,7	0,009
Asetelma 2 b	2	4,7	0,010
Asetelma 3	2	4,7	0,009

Kun tutkittiin lähemmin sukupuolen ja vyötärönympäryksen yhdysvaikutusta huomattiin, että vertailuasetelma 0 antoi jälleen liberaaleimmat testitulokset (taulukko 6.5.). Asetelma 1 (PML-menetelmä), asetelmat 2a ja 2b (GEE-menetelmän vaihdettavan korrelaatiorakenteen versio) sekä asetelma 3 (sekamalli) antoivat hyvin lähellä toisiaan olevat tulokset. Huomattakoon, että asetelmassa 1 ryvästymiseen reagoitiin vain beta-kerroinestimaattorivektorin kovarianssimatriisin estimoinnissa, mutta asetelmissa 2a, 2b ja 3 ryvästymiseen reagoitiin myös itse kerroinvektorin estimoinnissa.

Taulukko 6.5.
Yhdysvaikutustermin SEX(miehet)*RCIRCUM(luokka 2) asetelmaperusteiset testitulokset eri asetelmavalinnoilla.

	Beta-estimaatti	Keski-virhe	Asetelma-kerroin	t-testi	p-arvo
Asetelma 0 sex(1).rcircum(2)	0,02	0,01	1,00	2,14	0,03
Asetelma 1 (SUDAAN/REGRESS) sex(1).rcircum(2)	0,02	0,01	1,06	2,02	0,04
Asetelma 2a (SUDAAN/REGRESS) sex(1).rcircum(2)	0,02	0,01	1,05	1,87	0,06
Asetelma 2b (SAS/GENMOD) sex(1).rcircum(2)	0,02	0,01	1,02	1,90	0,06
Asetelma3 (SAS/MIXED) sex(1).rcircum(2)	0,02	0,01	1,02	1,93	0,05

Pitkäaikaissairastavuus

Pitkäaikaissairastavuuden mallintamisessa käytettiin logistista ANCOVA-mallia, joka sovitettiin samojen analyysiasetelmien vallitessa kuin edellä. Selittäviä muuttujia olivat koulutus (vuosina) sekä ikä ja sukupuoli. Koulutusmuuttujana käytettiin kolmiluokkaista muuttujaa REDUC, joka muodostettiin samaan tapaan kuin muuttuja RCIRCUM edellisessä esimerkissä. Analyysiasetelma on taulukossa 6.6. Siinä havainnollisuuden vuoksi ikä on kolmiluokkaisena (analyysin yhteydessä ikää käytettiin jatkuvana muuttujana). Taulukon mukaan pitkäaikaissairastavuuden esiintyvyys vaihteli selvästi selittäjien muodostaman luokituksen mukaan. Taso nousi ikävuosien lisääntyessä, mutta laski koulutusajan kasvaessa kussakin ikäluokassa. Estimoidut asetelmakertoimet olivat yleensä jonkin verran ykköstä suurempia.

Taulukko 6.6.
Pitkäaikaissairastavuus sukupuolen, iän ja koulutuksen mukaan

Miehet ikäluokan ja koulutusluokan mukaan				
	n	**Keskiarvo**	**Keskivirhe**	**Asetelmakerroin**
30–45 vuotta				
1 (alin)	36	0,35	0,09	1,22
2	487	0,35	0,02	1,26
3 (ylin)	498	0,25	0,02	0,91
46–64 vuotta				
1 (alin)	393	0,66	0,03	1,23
2	475	0,51	0,02	1,12
3 (ylin)	337	0,45	0,03	1,01
65 vuotta				
1 (alin)	355	0,84	0,02	1,20
2	73	0,71	0,06	1,07
3 (ylin)	48	0,69	0,07	0,99
Kaikki	2 702	0,50	0,01	1,47
Naiset ikäluokan ja koulutusluokan mukaan				
	n	**Keskiarvo**	**Keskivirhe**	**Asetelmakerroin**
30–45 vuotta				
1 (alin)	36	0,59	0,09	1,28
2	389	0,38	0,02	0,91
3 (ylin)	733	0,30	0,02	1,03
46–64 vuotta				
1 (alin)	331	0,67	0,03	1,27
2	594	0,55	0,02	0,82
3 (ylin)	370	0,45	0,03	0,96
65– vuotta				
1 (alin)	551	0,82	0,02	1,53
2	139	0,78	0,04	0,99
3 (ylin)	71	0,70	0,05	0,94
Kaikki	3 214	0,53	0,01	1,19

Tarkastellaan lähemmin analyysituloksia vertailuasetelmilla 0 sekä asetelmaperusteisilla asetelmilla 2a ja 2b (täydellisemmät tulokset on esitetty liitetaulukossa 2). Tulosten mukaan vertailuasetelma 0 (SRS-oletus, ei painoja) tuotti konservatiivisimmat tulokset. Asetelmat 2a ja 2b (SUDAAN/RLOGIST tai SAS/GENMOD) tuottivat toisiaan lähellä olevat tulokset. Esimerkiksi sukupuolen vaikutusta kuvaavalle termille saatiin taulukon 6.7. mukaiset tulokset.

Taulukko 6.7.
Sukupuolen vaikutusta kuvaavan asetelmaperusteisen t-testin tulokset eri analyysiasetelmilla

	Beta-estimaatti	Keski-virhe	Asetelma-kerroin	t-testi	p-arvo
Asetelma 0					
SEX	0,10	0,06	1,00	1,83	0,07
Asetelma 2a (SUDAAN/REGRESS)					
SEX	0,11	0,05	0,91	2,08	0,04
Asetelma 2b (SAS/GENMOD)					
SEX	0,11	0,05	0,87	2,12	0,03

6.4 Yhteenveto ja suosituksia

Terveys 2000 -tutkimuksen otanta-asetelma oli yhdistelmä henkilötason ositetusta otannasta ja kaksiasteisesta ryväsotannasta, jossa rypäinä olivat terveyskeskuspiirit. Esimerkkianalyyseissa käytettiin aineistoa, jossa oli 5 954 henkilöä. Aineistoon muodostettiin ensin asetelmapainot (sisältymistodennäköisyyden käänteisluku), joita uudelleenpainotuksen yhteydessä muokattiin niin, että vastauskadon vaikutusta otettiin huomioon. Lopulliset analyysipainot skaalattiin niin, että niiden keskiarvo aineistossa oli yksi.

Terveys 2000 -aineiston monimuuttuja-analyysin yhteydessä huomioon otettavia asetelman ominaisuuksia ovat:

- Ositus (20 alueellista ositetta, joista 15 ositteessa henkilötasoinen otanta ja viidessä kaksiasteinen otanta rypäinä terveyskeskuspiirit).

- Ryvästymisen (65 otosryvästä asetelman kaksiasteisessa osassa) aiheuttama havaintojen positiivinen sisäkorreloituneisuus. Tulosmuuttujan systolinen verenpaine keskiarvon estimoitu asetelmakerroin oli deff = 2.41 ja pitkäaikaissairastavuuden esiintyvyydelle deff = 1.75.

- Painotus (analyysipainot, jotka vaihtelevat ykkösen ympäristössä)

Otanta-asetelman ominaisuuksien huomioon ottamista varten lineaaristen ja logististen ANCOVA-mallien sovittamisen yhteydessä muodostettiin kolme varsinaista analyysiasetelmaa sekä vertailuasetelma 0, jossa kaikki otanta-asetelman ominaisuudet jätettiin pois analyysista. Lisäksi analyysiasetelma 2 jaettiin kahteen osaan (2a ja 2b) analyysin teknistä toteuttamista varten asetelmaperusteisella analyysiohjelmalla (SUDAAN/REGRESS ja LOGISTIC/RLOGIST) ja malliperusteisella analyysiohjelmalla (SAS/GENMOD). Asetelmien olennaiset piirteet ja erot olivat:

Asetelma 0 on vertailuasetelma, jossa oletettiin otos poimituksi yksinkertaisella satunnaisotannalla palauttaen (SRSWR). Samalla oletettiin, että havainnot ovat toisistaan riippumattomia. Keskivirheet estimoitiin malliperusteisesti. Painot olivat tässä tapauksessa ykkösiä. Analyysi voidaan tehdä millä tahansa tavanomaisen tilastollisen ohjelmiston ohjelmalla.

Asetelma 1 edustaa perinteistä asetelmaperusteista analyysiasetelmaa, jolla voidaan reagoida asetelman kaikkiin ominaisuuksiin (SUDAAN-ohjelmisto). Estimointi tehtiin painotetulla PNS-menetelmällä (lineaarinen malli) tai PML-menetelmällä (logistinen malli). Keskivirheiden estimointi tehtiin asetelmaperusteisesti (sandwich-varianssiestimaattori) käyttäen analyysipainoja. Tässä asetelmassa ryvästymiseen reagoitiin keskivirheiden estimoinnin yhteydessä, mutta ei mallin beta-parametrien estimoinnissa.

Asetelma 2 on asetelmaperusteinen analyysiasetelma, jolla voitiin reagoida asetelman ominaisuuksiin niin, että asetelma 2a kattoi kaikki ominaisuudet (SUDAAN) ja asetelma 2b muut paitsi osituksen (SAS/GENMOD). Estimointi tehtiin painotetulla GEE-menetelmällä, jossa käytettiin vaihdettavan korrelaatiorakenteen oletusta rypäiden sisällä. Keskivirheet estimoitiin asetelmaperusteisesti (sandwich-varianssiestimaattori) käyttäen analyysipainoja. Tässä asetelmassa ryvästymiseen reagoitiin sekä keskivirheiden estimoinnin yhteydessä että mallin beta-parametrien estimoinnissa. Analyysiasetelmassa tehdään voimakkaampia oletuksia kuin asetelmassa 1 (vakiokorrelaatio-oletus rypäiden sisällä), ja se voi tuottaa tehokkuushyötyä estimoinnissa asetelmaan 1 verrattuna.

Asetelma 3 on malliperusteinen analyysiasetelma, jossa käytettiin sekamalleja (SAS/MIXED). Ryvästymiseen reagoitiin parametrisoimalla malliin ryväskohtaiset satunnaistermit. Kiinteiden tekijöiden estimointi tehtiin GLS-menetelmällä ja satunnaistermien estimointi REML-menetelmällä. Keskivirheet estimoitiin sandwich-varianssiestimaattorin avulla. Tässä asetelmassa ryvästymiseen reagoitiin sekä keskivirheiden estimoinnin yhteydessä että mallin beta-parametrien estimoinnissa. Analyysipainot olivat mukana, mutta asetelman ositukseen ei reagoitu.

Tulokset osoittavat, että vertailuasetelman 0 mukaisilla analyyseilla on taipumus tuottaa liian moniparametrisia malleja erityisesti, jos tarkasteltavat tulosmuuttujat ovat positiivisesti sisäkorreloituneita otosrypäissä. Tällöin keskivirhe-estimaatit saattavat olla liian pieniä ja vastaavasti t-testisuureiden havaitut arvot liian suuria. Tätä analyysiasetelmaa ei voida suositella käytettäväksi päämenetelmänä. Suositeltavampia ovat analyysiasetelmat 1, 2 ja 3, joiden avulla voidaan ottaa huomioon asetelman ominaisuuksia.

Käytännön kannalta tärkeimmäksi osoittautui asetelmaan sisältyvän ryvästymisen sisällyttäminen analyysiin. Tämä voidaan tehdä asetelmaperusteisilla tai malliperusteisilla tekniikoilla ja ohjelmistoilla, joihin sisältyy mahdolli-

suus parametrisoida rypäänsisäinen korrelaatiorakenne estimointiyhtälöissä ("working" korrelaatio GEE-menetelmän vaihdettavan korrelaatiorakenteen versiossa) tai mallin rakenteessa (ryväskohtaiset satunnaistermit sekamalleissa) sekä analyysipainojen mukaanotto. GEE-menetelmän asetelmaperusteiset versiot ovat käytettävissä SUDAAN-ohjelmistossa (REGRESS, LOGISTIC/RLOGIST ja MULTILOG). Tällöin on mahdollista sisällyttää analyysiin kaikki asetelman piirteet eli ositus, ryvästys ja painotus. GEE-menetelmä on käytettävissä myös SAS-ohjelmiston proseduurissa GENMOD, mutta osituksen kontribuutiota on vaikea hallita analyysissa. Selkeimmin malliperusteista analyysia edustaa sekamalliin perustuva asetelma 3 (SAS/MIXED).

Analyyseissa havaittiin, että analyysiasetelmilla 1, 2 ja 3 päädyttiin varsin yhtäpitäviin numeerisiin tuloksiin piste-estimaattien ja keskivirhe-estimaattien sekä t-testisuureiden osalta. Asetelman 1 ero asetelmiin 2 ja 3 nähden on, että asetelmaperusteisessa asetelmassa 1 toimitaan lievemmillä olettamuksilla kuin asetelmissa 2 ja 3. Analyysiasetelma 1 edustaa siten selkeimmin robustia asetelmaperusteista menetelmää ja on turvallinen perusvalinta.

Hyvään analyysistrategiaan kuuluu, että monimuuttuja-analyysi tehdään eksploratiivisesti eri malliolettamusten vallitessa ja vaihtoehtoisilla menetelmävalinnoilla. Näin voidaan tutkittavasta ilmiöstä saada riittävän monipuolinen kuva luotettavia tilastollisia päätelmiä varten.

Lähteet

AAPOR (2000): *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys.* The American Association for Public Opinion Research. Ann Arbor, Michigan.

Aromaa, A., Heliövaara, M., Impivaara, O., Knekt, P. & Maatela, J. (1989). Tutkimuskohteet, toteutus ja aineisto. Teoksessa: Aromaa, A., Heliövaara, M., Impivaara, O., Knekt, P., Maatela, J., Joukamaa, M., Klaukka, T., Lehtinen, V., Melkas, T., Mälkiä, E., Nyman, K., Paunio, I., Reunanen, A., Sievers, K., Kalimo, E. & Kallio, V. *Terveys, toimintakyky ja hoidontarve Suomessa. Mini-Suomi-terveystutkimuksen perustulokset.* Kansaneläkelaitoksen julkaisuja AL: 32. Helsinki ja Turku: Kansaneläkelaitos. 33–62.

Aromaa, A. & Koskinen, S. (2002). Aineisto ja menetelmät. Teoksessa: Aromaa, A. & Koskinen, S. (toim.). *Terveys ja toimintakyky Suomessa. Terveys 2000 -tutkimuksen perustulokset.* Kansanterveyslaitoksen julkaisuja B3/2002. Helsinki: Kansanterveyslaitos. 3–15.

Cochran, W. G. (1963). *Sampling Techniques.* 2nd ed. New York: John Wiley & Sons.

Deming, W. E. (1953). On a Probability Mechanism to Attain an Economic Balance Between the Resultant Error of Response and the Bias of Nonresponse. *Journal of the American Statistical Association.* 48. 264. 743–772.

Diggle, P. J., Liang, K.-Y. & Zeger, S. L. (1994). *Analysis of Longitudinal Data.* Oxford: Oxford University Press.

Djerf, K. (2000). *Properties of Some Estimators under Unit Nonresponse.* Statistics Finland. Research Reports 231. Helsinki: Statistics Finland.

Ekholm, A. & Laaksonen, S. (1991). Weighting via Response Modeling in the Finnish Household Budget Survey. *Journal of Official Statistics.* Vol. 7. No. 3. 325–337.

Glenn, N. D. (1969). Aging, Disengagement, and Opinionation. *Public Opinion Quarterly.* Vol. 33. 17–33.

Goldstein, H. (1995). *Multilevel Statistical Models.* 2nd edition. London: Arnold; New York: John Wiley & Sons.

Graubard, B. I. & Korn, E. L. (1999). Predictive Margins with Survey Data. *Biometrics.* Vol. 55. No. 2. 652–659.

Groves, R. M., Cialdini, R. B. & Couper, M. P. (1992). Understanding the Decision to Participate in a Survey. *Public Opinion Quarterly.* Vol. 56. No 4. 475–495.

Groves, R. M. & Couper, M. P. (1998). *Nonresponse in Household Interview Surveys.* New York: John Wiley & Sons.

Hansen, M. H. & Hurwitz, W. N. (1946). The problem of nonresponse in sample surveys. Journal of the American Statistical Association. Vol. 41. 517–529.

Hansen, M. H., Hurwitz, W. N. & Madow, W. G. (1953). *Sample Survey Methods and Theory*. New York: John Wiley & Sons.

Horton, N. J. & Lipsitz, S. R. (1999). Review of Software to Fit Generalized Estimating Equation Regression Models. *The American Statistician*. Vol. 53. No. 2. 160–169.

Kattainen, A., Koskinen, S., Reunanen, A., Martelin, T., Knekt, P. & Aromaa, A. (2003). Impact of cardiovascular diseases on activity limitations and need for help in elderly Finns. *Journal of Clinical Epidemiology*. (In print).

Kemsley, W. F. F. (1975). Family Expenditure Survey. A Study of Differential Response Based on a Comparison of the 1971 Sample with Census. *Statistical News*. Vol. 35. 18–22.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons.

Kish, L. (1987). *Statistical Design for Research*. New York: John Wiley & Sons.

Koponen, P. & Aromaa, A. (2003). *Survey Design and Methodology in National Health Interview and Health Examination Surveys. Review of literature, European survey experiences and recommendations*. Health Surveys in the EU: HIS and HIS/HES Evaluations and Models. Phase 2/Subproject 3. European Comission.

Kuusela, V. & Parviainen, A. (1997). An Object-Based Case Management System for CAPI Surveys. Teoksessa: Boyeldieu I. (toim.): des Utilisateurs de BLAISE. INSEE: Paris. http://www.blaiseusers.org/Ibucpdfs/1995-1998/kuusel97.pdf

Kuusela, V., Tanskanen, V. & Virtala, E. (2000). Data Collection in Two Phase Multicenter Survey. 6th International Blaise Users' Confrence. Kinsale, Ireland. http://www.blaiseusers.org/Ibucpdfs/2000/vesa_multi_centre_health_survey.pdf

Kuusela, V. (2001): A Windows based User Interface and CAPI Information System. 7th International Blaise Users' Confrence. Washington DC, USA. http://www.blaiseusers.org/ibucpdfs/2001/Kuusela-IBUC_paper.pdf

Kviz, F. J. (1977). Toward a standard definition of response rate. Public Opinion Quarterly. Vol. 41. 265–267.

Laiho, J. (2001). *Modelling Interviewer Effects and Survey Participation Processes*. 12th International Workshop on Household Non-Response. Oslo, Norway.

Laiho, J. (2002a). Vastausosuuksien merkitys otantatutkimusten vertailukelpoisuuden näkökulmasta. *Hyvinvointikatsaus* 2/2002. Tilastokeskus.

Laiho, J. (2002b). *Estimating the Impact of Survey Errors on Survey Estimates.* The International Conference on Improving Surveys (ICIS). Conference CD. Copenhagen, Denmark. www.icis.dk/ICIS_papers/E3_3_3.pdf

Laiho, J. & Hietaniemi, L. (toim.). (2002). *Laatua tilastoissa.* Käsikirjoja 43. Helsinki: Tilastokeskus.

Lee, J. (1981). Covariance Adjustment of Rates Based on the Multiple Logistic Regression Model. *Journal of Chronic Diseases.* Vol. 34. 415–426.

Lehtonen, R. (1996). Interviewer Attitudes and Unit Nonresponse in Two Different Interviewing Schemes. Teoksessa: Laaksonen, S. (toim.). *International Perspectives on Nonresponse. Proceedings of the Sixth International Workshop on Household Survey Nonresponse.* Research Reports 219. Helsinki: Statistics Finland.130–140.

Lehtonen, R. & Pahkinen, E. (1996). Practical Methods for Design and Analysis of Complex Surveys. Revised Edition. Chichester: John Wiley & Sons, Ltd. (Second Edition 2004).

Lehtonen, R., Djerf, K., Härkänen, T. & Laiho, J. (2003a). A Comparison of Design-Based and Model-Based Methods for the Analysis of Complex Health Survey Data: A Case Study. Ottawa: Proceedings of Statistics Canada Methodology Symposium 2002, *Modelling Survey Data for Social and Economic Research.*

Lehtonen, R., Djerf, K., Härkänen, T. & Laiho, J. (2003b). Modelling Complex Health Survey Data: A Case Study. Teoksessa: Höglund, R. Jäntti, M., & Rosenqvist, G. (toim.). Statistics, Econometrics and Society: Essays in Honour of Leif Nordberg. Helsinki: Statistics Finland, Research Reports 238, 91-114.

Liang, K.-Y. & Zeger, S. L. (1986). Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika.* Vol. 73. Issue 1. 13–22.

Lindström, H. L. (1983). *Non-Response Errors in Sample Surveys.* Urval. No. 16. Örebro: Statistics Sweden.

Lohr, S. L. (1999). *Sampling: Design and Analysis.* Pacific Grove: Duxbury Press.

Lynn, P., Beerten, R., Laiho, J. & Martin, J. (2002). Towards standardisation of survey outcome categories and response rate calculations. *Research in Official Statistics.* Vol. 5. No. 1. 61–84.

McCulloch, C. E. & Searle, S. R. (2001). *Generalized, Linear, and Mixed Models.* New York: John Wiley & Sons.

Nieminen, M. (2003). *Väestötietojärjestelmän luotettavuustutkimus 2002.* Raportti 31.1.2003. Elinolot/Haastattelu- ja tutkimuspalvelut. Tilastokeskus.

Nieminen, T. (1997). *Terveydenhoitajat ja ammattihaastattelijat terveystietojen kerääjinä. Vertaileva tutkimus Terveydenhuollon väestötutkimuksen 1995-1996 aineiston keruumenetelmistä.* Sosiaali- ja terveysturvan tutkimuksia 27. Helsinki: Kansaneläkelaitos.

Nieminen, T. (2003). Haastattelijapalautetta Terveys 2000 -tutkimuksesta. *Hyvinvointikatsaus* 2/2003. Tilastokeskus.

Oh, J. L. & Scheuren, F. (1983). Weighting Adjustment for Unit Nonresponse. Teoksessa: Madow, W. G., Olkin, I. & Rubin, D. B. (toim.). *Incomplete Data in Sample Surveys.* Vol. 2. New York: Academic Press. 143–184.

Pfeffermann, D., Skinner, C. J., Goldstein, H., Holmes, D. J. & Rasbash, J. (1998). Weighting for Unequal Selection Probabilities in Multilevel Models (with discussion). *Journal of the Royal Statistical Society.* Series B. Vol. 60. Issue 1. 23–40.

Platek, R. & Gray, G.B. (1986) On the Definitions of Response Rates. *Survey Methodology.* Vol. 12. 17–27.

Politz, A. & Simmons, W. (1949). An attempt to Get the "Not-at-Homes" into the Sample WIthout Call-Backs. *Journal of the American Association.* Vol. 44. 9–31.

Research Triangle Institute (2001). *SUDAAN User's Manual, Release 8.0.* Research Triangle Park, NC: Research Triangle Institute.

Ruotsalainen, K. (2002). Use of Administrative Data in Population Censuses – Definition of Main Type of Activity as an Example. *UNECE Works session on Statistical Editing.* Helsinki, Finland.

Sautory, O. (1993). *La macro CALMAR. Redressement d'un échantillon par calage sur marges.* I.N.S.E.E. Série des documents de travail *n°* F 9310. Paris: I.N.S.E.E.

Singer, J. D. (1998). Using SAS PROC MIXED to Fit Multilevel Models, Hierarchical Models, and Individual Growth Models. *Journal of Educational and Behavioral Statistics.* Vol. 24. No. 4. 323–355.

Skinner, C., Holt, T. & Smith, T. M. F. (toim.). (1989). *Analysis of Complex Surveys.* Chichester: John Wiley & Sons.

Smith, T. W. (1983). The Hidden 25 Percent: An Analysis of Nonresponse on the 1980 General Social Survey. *Public Opinion Quarterly.* Vol. 47. 386–404.

Smith, T. W. (2002). Developing Nonresponse Standards. Teoksessa: Groves, R.M., Dillman, D. A., Eltinge, J. L., Little, R. J. A. (toim.) *Survey Nonresponse.* New York: John Wiley & Sons. 27–40.

Särndal, C.-E., Swensson, B. & Wretman, J. (1992). *Model Assisted Survey Sampling.* New York: Springer-Verlag.

Tilastokeskus. (2001). *Väestölaskenta 2000.* Käsikirjoja 35. Helsinki: Tilastokeskus.

Tourangeau, R., Rips, L.J. & Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge, England: Cambridge University Press.

Ylitalo, M. (2002). *Väestötietojärjestelmän luotettavuustutkimus 2001*. Raportti 15.2.2002. Elinolot/Haastattelu- ja tutkimuspalvelut. Tilastokeskus.

Ziegler, A., Kastner, C. & Blettner, M. (1998). The Generalized Estimating Equations: An Annotated Bibliography. *Biometrical Journal*. Vol. 40. Issue 2. 115–139.

Liitteet

Liite 1.

Terveys 2000 -tutkimukseen poimitut terveyskeskuspiirit ja kohdehenkilöiden poimintavälit ositteen ja miljoonapiirin mukaan

Koodi		Poimintavälit	
Itse-edustavat ositteet:	**Terveyskeskuspiiri**	**18–79-vuotiaat**	**yli 80- vuotiaat**
124	Espoo	397	198
152	Helsinki	398	199
153	Vantaa	397	198
270	Kotka	399	199
305	Lappeenranta	399	199
545	Turku	398	199
424	Pori	397	198
635	Hämeenlinna	398	199
532	Tampere	398	199
299	Lahti	399	199
565	Vaasa	396	198
181	Joensuu	398	199
280	Kuopio	398	199
192	Jyväskylä	397	198
390	Oulu	397	198
Muut rypäät: 1 HYKS			
162	Hyvinkää	293	146
219	Karkkila	61	30
426	Porvoo	306	153
550	Tuusula	200	100
696	Lohja	297	148
637	Loviisa	137	68
709	Hamina	190	95
700	Kouvola	311	155
176	Imatra	229	114
338	Luumäki	39	19
470	Ruokolahti	46	23
2 TYKS			
392	Parainen	153	76
631	Somero	129	64
686	Uusikaupunki	227	113
697	Kaarina	322	161
703	Loimaa	251	125
657	Masku	163	81
648	Naantali	205	102
623	Perniö	92	46
719	Salo	589	294
728	Vehmaa	71	35
258	Kokemäki	117	58
688	Harjavalta	233	116
677	Rauma	622	311
716	Ulvila	180	90

Liite 1. jatkuu

Terveys 2000 -tutkimukseen poimitut terveyskeskuspiirit ja kohdehenkilöiden poimintavälit ositteen ja miljoonapiirin mukaan

Koodi		Poimintavälit	
Itse-edustavat ositteet:	Terveyskeskuspiiri	18–79-vuotiaat	yli 80-vuotiaat
3 TaYKS			
614	Forssa	228	57
726	Riihimäki	256	128
375	Nokia	162	81
567	Valkeakoski	133	66
615	Heinola	285	142
624	Orimattila	119	59
169	Ilmajoki	71	35
308	Lapua	87	43
683	Seinäjoki	306	153
659	Uusikaarlepyy	46	23
712	Kristiinankaupunki	82	41
676	Pietarsaari	198	99
4 KYS			
674	Mikkeli	439	219
706	Savonlinna	304	152
620	Kerimäki	76	38
662	Pyhäselkä	70	35
189	Juuka	54	27
612	Lieksa	128	64
422	Polvijärvi	43	21
632	Siilinjärvi	175	87
303	Lapinlahti	61	30
682	Iisalmi	222	111
707	Jämsä	191	95
651	Keuruu	114	57
670	Muurame	97	48
5 OYS			
634	Kokkola	439	219
145	Haukipudas	140	70
247	Kiiminki	86	43
286	Kuusamo	173	36
374	Nivala	104	52
440	Pyhäjärvi	71	35
527	Taivalkoski	50	25
643	Raahe	324	162
711	Simo	61	30
627	Ylivieska	200	100
203	Kajaani	366	183
232	Kemi	251	125
468	Rovaniemi	364	182
501	Sodankylä	102	51
559	Utsjoki	14	7

Liite 2.
Laatuselvityksessä käytetyt käsitteet, määritelmät ja luokitukset

Demografiset käsitteet

Asuntokunta
Asuntokunta koostuu samassa asuinhuoneistossa vakinaisesti asuvista henkilöistä.

Kotitalous
Kotitalouden muodostavat henkilöt, joilla on kokonaan tai osittain yhteinen ruokatalous tai jotka muuten käyttävät tulojaan yhdessä. Kotitalous on laajempi käsite kuin perhe, mutta yleensä suppeampi kuin asuntokunta.

Kotitalouden määrittäminen ei ole aina suoraviivaista, minkä vuoksi kotitalousmääritelmään on tehty joitain yleisiä rajanvetoja. Esimerkiksi ase- tai siviilipalvelustaan suorittavat, toisella paikkakunnalla tai ulkomailla tilapäisesti työssä olevat, sairaalahoidossa tilapäisesti ja lomalla tai matkoilla olevat luetaan aina kotitaloutensa jäseniksi. Toisella paikkakunnalla opiskelevat koululaiset ja opiskelijat luetaan vanhempiensa kotitalouteen, jos he elävät pääosin vanhempiensa tuloilla. Jos he elävät omilla tuloillaan kuten opintotuella, he muodostavat oman kotitaloutensa. (Laiho, 1998).

Terveys 2000 -tutkimuksen terveyshaastattelulomakkeessa haastateltaville selitettiin kotitalouden käsite seuraavasti:

'Tässä haastattelussa tarkoitamme kotitaloudella henkilöitä, jotka asuvat ja ruokailevat yhdessä tai jotka muutoin käyttävät tulojaan yhdessä.'

Kotitalousväestö
Kaikki yhden tai useamman hengen kotitalouksissa elävät henkilöt muodostavat yhdessä kotitalousväestön.

Laitosväestö
Laitoksissa vakituisesti olevat henkilöt muodostavat laitosväestön.

Tiedonkeruumenetelmien keskeisiä käsitteitä

CAPI
Tietokoneavusteinen henkilöhaastattelu. (*Computer Assisted Personal Interview*).

PAPI
Paperilomakeavusteinen henkilöhaastattelu. (*Paper-and-Pencil Interview*).

Otantatutkimuksen keskeisiä käsitteitä

Tavoiteperusjoukko
Tavoiteperusjoukko on kiinteä ja äärellinen perusjoukko, johon tutkimus tahdotaan kohdistaa.

Kohdeperusjoukko
Kohdeperusjoukko on kiinteä eli äärellinen perusjoukko, johon tutkimus kyetään otoskehikon puitteissa kohdistamaan.

Kehikkoperusjoukko
Kehikkoperusjoukko pyritään muodostamaan mahdollisimman samanlaiseksi kuin kohdeperusjoukko.

Otoskehikko
Otoskehikko on rekisteri tai tietokanta, joka muodostaa kehikkoperusjoukon ja josta otos poimitaan.

Ylipeitto
Ylipeitto koostuu otoskehikon alkioista (tässä tapauksessa henkilöistä), jotka ovat joko lakanneet olemasta (kuolleet, muuttaneet maasta tai poimitusta terveyskeskuspiiristä) tai otoskehikon ylläpitäjän käytettävissä olevan informaation riittämättömyyden vuoksi jääneet otoskehikkoon.

Alipeitto
Alipeitto koostuu alkioista, joita ei eri syistä ole ollut otoskehikossa mukana. Alipeiton merkittävyyttä on vaikeampi arvioida kuin ylipeiton.

Kohdehenkilö
Henkilöotokseen poimittuja henkilöitä kutsutaan kohdehenkilöiksi.

Kato
Ne kohdehenkilöt, joita ei ole pystytty eri syistä haastatella tutkimukseen. Kadon syyt luokitellaan kieltäytymisiin, tavoitettavuus vaikeuksiin, ja muihin syihin kuten esimerkiksi kielivaikeuksiin ja sairaalassaolojaksoihin.

Sisältymistodennäköisyys
Satunnaisotantaan perustuvissa otantamenetelmissä käytetty otanta-asetelma määrittää kohdehenkilöiden todennäköisyyden tulla poimituksi otokseen.

Painokerroin
Painokerroin korottaa estimaattorit perusjoukon eli korjaa kohdehenkilöistä estimoidut jakaumat vastaamaan koko kohdeperusjoukon jakaumia. Painokertoimet korjaavat eri väestöryhmien osuudet otoksessa vastaamaan väestön vastaavia osuuksia. Yksinkertainen painokerroin on sisältymistodennäköisyyden käänteisluku. Korottavan painokertoimen avulla voidaan saada perusjoukkoa kuvaavia satunnaislukuja (esimerkiksi esiintymisfrekvenssejä tai muuttujan kokonaismääriä).

Kalibrointi
Kalibrointi on painokertoimien muodostukseen liittyvä tilastollinen menettely, jossa määritetään otokselle vastaavat väestön reunajakaumat. Kalibroinnin lähtötietona käytetään yksinkertaista korotuskerrointa, joka on sisältymistodennäköisyyden käänteisluku. Kalibrointi proseduurissa tätä painokerrointa pyritään pakottamaan interoivien vaiheiden kautta siten, että alkuperäisen korotuskertoimen muutoksen minimoidaan ja saman aikaisesti kalibroitu painokerroin tuottaa oikean väestöjakauman mukaiset reunajakaumat.

Analyysipaino
Analyysipainolla tarkoitetaan lopullista skaalattua (ja kalibroitua) painokerrointa, jonka keskiarvo on 1 ja summa vastanneiden lukumäärä. Analyysipaino ottaa siis otanta-asetelman ja siihen kalibroimalla tehdyt muokkaukset huomioon.

Keskeiset raportissa käytetyt luokitukset

Alueluokitukset:

Terveyskeskuspiiri
Terveyskeskuspiirillä tarkoitetaan sitä kuntaa tai kuntayhtymää, joka ylläpitää terveyskeskusta.

Miljoonapiiri eli yliopistosairaalapiiri
Miljoonapiirit määräytyvät yliopistollisten keskussairaaloiden mukaan, joita Suomessa on viisi:

- Helsingin yliopistollinen keskussairaalapiiri (HYKS),
- Turun yliopistollinen keskussairaalapiiri (TYKS),
- Tampereen yliopistollinen keskussairaalapiiri (TaYS),
- Kuopion yliopistollinen keskussairaalapiiri (KYS) ja
- Oulun yliopistollinen keskussairaalapiiri (OYS).

Koulutusluokitus

Koulutusaste
Kohdehenkilön koulutusaste on muodostettu Tilastokeskuksen tutkintorekisteristä. Koulutusluokitus on tarkoitettu mittaamaan koulujärjestelmäkoulutusta eli peruskouluissa, lukioissa, ammatillisissa oppilaitoksissa, ammattikorkeakouluissa ja yliopistoissa annettavaa pidempikestoista, pääsääntöisesti kokopäivätoimisesti järjestettyä tutkintoon tai koulutusammattiin tähtäävää koulutusta. Koulutusluokituksen pääkriteerit ovat koulutusaste ja koulutusala. Tässä laatuselvityksessä on käytetty koulutusluokituksen 1-numero tasoa:

- perusaste tai ei koulutusta,
- alempi keskiaste,
- ylempi keskiaste,
- alin korkea-aste,
- alempi kandidaattiaste,
- ylempi kandidaattiaste,
- tutkijakoulutus tai vastaava ja
- koulutusaste tuntematon.

Sosioekonomisen aseman ja tulojen luokitus

Sosioekonominen asema

Tässä laatuselvityksessä on käytetty hyvin karkeata pääasiallisen sosioekonomisen aseman luokitusta, joka on pystytty johtamaan verottajan rekisteristä. Luokitus perustuu Tulonjakotilaston otanta-asetelmassa käytettyyn poimintaryhmien ositukseen. Pääasiallisen sosioekonomisen aseman luokitus on seuraava:

- palkansaajat,
- yrittäjät,
- maatalousyrittäjät,
- eläkeläiset ja
- muut.

Tässä esitetty luokitus ei ole virallinen sosioekonomisen aseman luokitus, mutta se on käyttökelpoinen tilanteessa, jossa käytettävissä on vain rekistereistä johdettua tietoa koko kohderyhmälle. Mikäli henkilön valtion veronalaisen palkkatulon ja pääomatulon summasta puolet on veronalaista palkkatuloa, yrittäjätuloa, maataloustuloa tai eläketuloa luokitellaan henkilö kyseen omaiseen ryhmään. Muussa tapauksessa henkilö ryhmitellään 'muut'-ryhmään.

Tulodesiilit

Desiiliryhmittäisessä analyysissa perusjoukko jaetaan tulojen suuruuden perusteella kymmeneen lukumäärältään yhtä suureen ryhmään. Ensimmäiseen desiiliryhmään tulee aineiston pienituloisin kymmenes ja kymmenenteen suurituloisin. Tämän laatuselvityksen tarkasteluissa on käytetty valtion verotuksen alaista tuloa ja tietolähteenä on verottajan rekisteri.

Liite 3.
Lineaaristen ANCOVA-mallien sovittaminen SUDAAN- ja SAS-ohjelmistoilla

Ohjelmakoodit

Analyysioptio 0 (vertailuasetelma).

```
proc regress data =health2000 filetype=sas design=srs
r=independent;
model SYSBP    = SEX AGE RCIRCUM SEX* RCIRCUM AGE* RCIRCUM;
```

Analyysiasetelma 1 (GEE-Independent).

```
proc regress data = health2000 filetype=sas design=wr
r=independent;
    nest STRATUM SUBJECT;
    weight REWEIGHT;
    model SYSBP  = SEX AGE RCIRCUM SEX* RCIRCUM AGE* RCIRCUM;
```

Analyysiasetelma 2a (GEE-Exchangeable).

```
proc regress data = health2000 filetype=sas design=wr
r=exchangeable;
    nest STRATUM SUBJECT;
    weight REWEIGHT;
    model SYSBP = SEX AGE RCIRCUM SEX* RCIRCUM AGE* RCIRCUM;
```

Analyysiasetelma 2b (GEE-Exchangeable).

```
proc genmod data = health2000;
    class SUBJECT;
    weight REWEIGHT;
    model SYSBP = SEX AGE RCIRCUM SEX* RCIRCUM
    AGE* RCIRCUM /error=normal;
    repeated subject=SUBJECT / type=exch;
```

Analyysiasetelma 3 (Lineaarinen sekamalli).

```
proc mixed data = health2000 empirical method=reml;
    class SUBJECT;
    weight REWEIGHT;
    model SYSBP = SEX AGE RCIRCUM SEX*RCIRCUM
    AGE*RCIRCUM /solution;
    repeated / subject=SUBJECT type=vc;
```

Muuttujaviittaukset:

Asetelmaindikaattorit:	STRATUM	Ositemuuttuja
	SUBJECT	Ryväsmuuttuja
	REWEIGHT	
	Painomuuttuja	
Tulosmuuttuja:	SYSBP	Systolinen verenpaine
Selittävät muuttujat:	SEX	Sukupuoli,
	AGE	Ikä (vuosina)
	RCIRCUM	Vyötärönympärys (kolmiluokkainen)

Liite 4.

Tulostusotteet.

Lineaarisen ANCOVA-mallin sovittaminen systoliselle verenpaineelle (logaritmimuunnos) selittäjinä sukupuoli (SEX), ikä vuosina (AGE) ja kolmiluokkainen vyötärönympärys (RCIRCUM).

Analyysiasetelma 0.
Kiinteiden tekijöiden lineaarinen ANCOVA-malli (SUDAAN proseduuri REGRESS). Yksinkertainen satunnaisotanta, LS-estimointi, mallipe-rusteiset keskivirheet, ei analyysipainoja (vertailuasetelma).

Parametrit		Beta-estimaatti	Keskivirhe	t-testi	p-arvo
Vakio		4,744	0,0135	352,6	0,000
AGE		0,004	0,0002	15,8	0,000
SEX					
1		-0,009	0,0062	-1,5	0,148
2		0,000	0,0000	.	.
RCIRCUM					
1		-0,206	0,0170	-12,1	0,000
2		-0,121	0,0185	-6,5	0,000
3		0,000	0,0000	.	.
AGE,	RCIRCUM				
1,	1	0,002	0,0003	7,1	0,000
1,	2	0,001	0,0003	4,2	0,000
1,	3	0,000	0,0000	.	.
SEX,	RCIRCUM				
1,	1	0,031	0,0094	3,2	0,001
1,	2	0,019	0,0088	2,1	0,032
1,	3	0,000	0,0000	.	.
2,	1	0,000	0,0000	.	.
2,	2	0,000	0,0000	.	.
2,	3	0,000	0,0000	.	.

Analyysiasetelma 1.

Kiinteiden tekijöiden lineaarinen ANCOVA-malli (SUDAAN proseduuri REGRESS). Ositettu ryväsotanta, WLS-estimointi, asetelmaperusteiset keskivirheet, analyysipainot.

Parametrit		Beta-estimaatti	Deff Beta	Keskivirhe	t-testi	p-arvo
Vakio		4,742	1,24	0,0149	317,4	0,0000
AGE		0,004	1,20	0,0002	14,6	0,0000
SEX						
1		-0,008	1,08	0,0064	-1,2	0,2106
2		0,000	.	0,0000	.	.
RCIRCUM						
1		-0,208	1,13	0,0180	-11,5	0,0000
2		-0,120	1,16	0,0199	-6,0	0,0000
3		0,000	.	0,0000	.	.
AGE,	RCIRCUM					
1,	1	0,002	1,09	0,0003	6,9	0,0000
1,	2	0,001	1,12	0,0003	3,9	0,0001
1,	3	0,000	.	0,0000	.	.
SEX,	RCIRCUM					
1,	1	0,029	0,99	0,0094	3,1	0,0018
1,	2	0,018	1,06	0,0091	2,0	0,0430
1,	3	0,000	.	0,0000	.	.
2,	1	0,000	.	0,0000	.	.
2,	2	0,000	.	0,0000	.	.
2,	3	0,000	.	0,0000	.	.

Analyysiasetelma 2a.

Kiinteiden tekijöiden lineaarinen ANCOVA-malli (SUDAAN proseduuri REGRESS). Ositettu ryväsotanta, GEE-estimointi, "exchangeable" korrelaatiorakenne, asetelmaperusteiset keskivirheet, analyysipainot.

Parametrit		Beta-estimaatti	Deff Beta	Keskivirhe	t-testi	p-arvo
Vakio		4,744	1,17	0,0145	326,4	0,000
AGE		0,004	1,14	0,0002	14,7	0,000
SEX						
1		-0,008	1,02	0,0063	-1,2	0,218
2		0,000	.	0,0000	.	.
RCIRCUM						
1		-0,208	1,13	0,0181	-11,5	0,000
2		-0,123	1,12	0,0196	-6,3	0,000
3		0,000	.	0,0000	.	.
AGE,	RCIRCUM					
1,	1	0,002	1,10	0,0003	6,9	0,000
1,	2	0,001	1,05	0,0003	4,2	0,000
1,	3	0,000	.	0,0000	.	.
SEX	RCIRCUM					
1,	1	0,029	1,03	0,0096	3,0	0,003
1,	2	0,017	1,06	0,0091	1,9	0,061
1,	3	0,000	.	0,0000	.	.
2,	1	0,000	.	0,0000	.	.
2,	2	0,000	.	0,0000	.	.
2,	3	0,000	.	0,0000	.	.

Analyysiasetelma 2b.

Kiinteiden tekijöiden lineaarinen ANCOVA-malli (SAS-proseduuri GENMOD). Ryväsotanta, GEE-estimointi, "exchangeable" korrelaatiorakenne, asetelmaperusteiset keskivirheet, analyysipainot.

Parametrit			Beta-estimaatti	Keskivirhe	97.5% luottamusväli		Z-testi	p-arvo
Vakio			4,736	0,0127	4,707	4,764	73,5	<,0001
AGE			0,004	0,0002	0,003	0,004	14,5	<,0001
SEX	2		0,008	0,0062	-0,006	0,022	1,3	0,2114
SEX	1		0,000	0,0000	0,000	0,000	.	.
RCIRCUM	1		-0,179	0,0172	-0,218	-0,141	-10,5	<,0001
RCIRCUM	2		-0,106	0,0155	-0,141	-0,072	-6,9	<,0001
RCIRCUM	3		0,000	0,0000	0,000	0,000	.	.
AGE*RCIRCUM	1		0,002	0,0003	0,002	0,003	7,0	<,0001
AGE*RCIRCUM	2		0,001	0,0003	0,001	0,002	4,3	<,0001
AGE*RCIRCUM	3		0,000	0,0000	0,000	0,000	.	.
SEX*RCIRCUM	2	1	-0,029	0,0097	-0,050	-0,007	-2,9	0,0034
SEX*RCIRCUM	2	2	-0,017	0,0089	-0,037	0,003	-1,9	0,0573
SEX*RCIRCUM	2	3	0,000	0,0000	0,000	0,000	.	.
SEX*RCIRCUM	1	1	0,000	0,0000	0,000	0.000	.	.
SEX*RCIRCUM	1	2	0,000	0,0000	0,000	0,000	.	.
SEX*RCIRCUM	1	3	0,000	0,0000	0,000	0,000	.	.

Analyysiasetelma 3.

Lineaarinen sekamalli (ANCOVA) (SAS-proseduuri MIXED). Ryväsotanta, REML-estimointi, empiiriset ("sandwich") keskivirheet, analyysipainot.

Parametrit	SEX	RCIRCUM	Beta-estimaatti	Keskivirhe	Df	t-testi	p-arvo
Vakio			4,744	0,0147	2484	322,7	<,0001
AGE			0,003	0,0002	3434	14,6	<,0001
SEX	1		-0,008	0,0062	3434	-1,3	0,209
SEX	2		0,000	.	.	.	.
RCIRCUM		1	-0,208	0,0179	3434	-11,6	<,0001
RCIRCUM		2	-0,123	0,0196	3434	-6,3	<,0001
RCIRCUM		3	0,000	.	.	.	.
AGE*RCIRCUM		1	0,002	0,0003	3434	7,0	<,0001
AGE*RCIRCUM		2	0,001	0,0003	3434	4,2	<,0001
AGE*RCIRCUM		3	0,000	.	.	.	.
SEX*RCIRCUM	1	1	0,029	0,0097	3434	2,9	0,003
SEX*RCIRCUM	1	2	0,017	0,0089	3434	1,9	0,053
SEX*RCIRCUM	1	3	0,000	.	.	.	.
SEX*RCIRCUM	2	1	0,000	.	.	.	.
SEX*RCIRCUM	2	2	0,000	.	.	.	.
SEX*RCIRCUM	2	3	0,000	.	.	.	.

Liite 5
Tulostusotteet

Logistisen ANCOVA-mallin sovittaminen pitkäaikaissairastavuudelle selittäjinä sukupuoli, ikä (vuosina) ja kolmiluokkainen koulutustaso (REDUC),

Analyysiasetelma 0.
Kiinteiden tekijöiden logistinen ANCOVA-malli (SUDAAN proseduuri LOGISTIC/RLOGIST). Yksinkertainen satunnaisotanta, ML-estimointi, mallipe-rusteiset keskivirheet, ei analyysipainoja (vertailuasetelma).

Parametrit	Beta-estimaatti	Keskivirhe	t-testi	p-arvo
Vakio	−2,78	0,128	−21,8	0,00
AGE	0,05	0,003	18,9	0,00
SEX	0,10	0,057	1,8	0,07
REDUC				
1	0,72	0,086	8,4	0,00
2	0,32	0,066	4,9	0,00
3	0,00	0,000	.	.

Analyysiasetelma 2a.
Kiinteiden tekijöiden logistinen ANCOVA-malli (SUDAAN proseduuri LOGISTIC/RLOGIST). Ositettu ryväsotanta, GEE-estimointi, "exchangeable" korrelaatiorakenne, asetelmaperusteiset keskivirheet, analyysipainot.

Parametrit	Beta-estimaatti	Deff Beta	Keskivirhe	t-testi	p-arvo
Vakio	−2,81	1,08	0,133	−21,1	0,00
AGE	0,05	1,09	0,003	18,0	0,00
SEX	0,11	0,91	0,052	2,1	0,04
REDUC					
1	0,72	0,87	0,080	9,1	0,00
2	0,31	1,03	0,068	4,6	0,00
3	0,00	.	0,000	.	.

Analyysiasetelma 2b.
Kiinteiden tekijöiden logistinen ANCOVA-malli (SAS-proseduuri GENMOD). Ryväsotanta, GEE-estimointi, "exchangeable" korrelaatiorakenne, asetelma-perusteiset keskivirheet, analyysipainot.

Parametrit		Beta-estimaatti	Keskivirhe	97.5 % luottamusväli		Z-testi	p-arvo
Vakio		−2,81	0,131	−3,105	−2,52	−21,5	<,0001
AGE		0,05	0,003	0,042	0,05	18,5	<,0001
SEX		0,11	0,053	−0,006	0,23	2,1	0,03
REDUC	1	0,72	0,080	0,543	0,90	9,1	<,0001
REDUC	2	0,31	0,067	0,160	0,46	4,6	<,0001
REDUC	3	0,00	0,000	0,000	0,00	.	.

Kuvio- ja taulukkoluettelo

TUTKIMUKSIA-SARJA
RESEARCH REPORTS SERIES

Tilastokeskus on julkaissut Tutkimuksia v. 1966 alkaen, v. 1990 lähtien ovat ilmestyneet seuraavat:

164. **Henry Takala,** Kunnat ja kuntainliitot kansantalouden tilinpidossa. Tammikuu 1990. 60 s.

165. **Jarmo Hyrkkö,** Palkansaajien ansiotasoindeksi 1985=100. Tammikuu 1990. 66 s.

166. **Pekka Rytkönen,** Siivouspalvelu, ympäristöhuolto ja pesulapalvelu 1980-luvulla. Tammikuu 1990. 70 s.

167. **Jukka Muukkonen,** Luonnonvaratilinpito kestävän kehityksen kuvaajana. 1990. 119 s.

168. **Juha-Pekka Ollila,** Tieliikenteen tavarankuljetus 1980-luvulla. Helmikuu 1990. 45 s.

169. **Tuovi Allén – Seppo Laaksonen – Päivi Keinänen – Seija Ilmakunnas,** Palkkaa työstä ja sukupuolesta. Huhtikuu 1990. 90 s.

170. **Ari Tyrkkö,** Asuinolotiedot väestölaskennassa ja kotitaloustiedustelussa. Huhtikuu 1990. 63 s.

171. **Hannu Isoaho – Osmo Kivinen – Risto Rinne,** Nuorten koulutus ja kotitausta. Toukokuu 1990. 115 s.

171b. **Hannu Isoaho – Osmo Kivinen – Risto Rinne,** Education and the family background of the young in Finland. 1990. 115 pp.

172. **Tapani Valkonen – Tuija Martelin – Arja Rimpelä,** Eriarvoisuus kuoleman edessä. Sosioekonomiset kuolleisuuserot Suomessa 1971–85. Kesäkuu 1990. 145 s.

173. **Jukka Muukkonen,** Sustainable development and natural resource accounting. August 1990. 96 pp.

174. **Iiris Niemi – Hannu Pääkkönen,** Time use changes in Finland in the 1980s. August 1990. 118 pp.

175. **Väinö Kannisto,** Mortality of the elderly in late 19th and early 20th century Finland. August 1990. 50 pp.

176. **Tapani Valkonen – Tuija Martelin – Arja Rimpelä,** Socio-economic mortality differences in Finland 1971–85. December 1990. 108 pp.

177. **Jaana Lähteenmaa – Lasse Siurala,** Nuoret ja muutos. Tammikuu 1991. 211 s.

178. **Tuomo Martikainen – Risto Yrjönen,** Vaalit, puolueet ja yhteiskunnan muutos. Maaliskuu 1991. 120 s.

179. **Seppo Laaksonen,** Comparative Adjustments for Missingness in Short-term Panels. April 1991. 74 pp.

180. **Ágnes Babarczy – István Harcsa – Hannu Pääkkönen,** Time use trends in Finland and in Hungary. April 1991. 72 pp.

181. **Timo Matala,** Asumisen tuki 1988. Kesäkuu 1991. 64 s.

182. **Iiris Niemi – Parsla Eglite – Algimantas Mitrikas – V.D. Patrushev – Hannu Pääkkönen,** Time Use in Finland, Latvia, Lithuania and Russia. July 1991. 80 pp.

183. **Iiris Niemi – Hannu Pääkkönen,** Vuotuinen ajankäyttö. Joulukuu 1992. 83 s.

183b. **Iiris Niemi – Hannu Pääkkönen – Veli Rajaniemi – Seppo Laaksonen – Jarmo Lauri,** Vuotuinen ajankäyttö. Ajankäyttötutkimuksen 1987–88 taulukot. Elokuu 1991. 116 s.

184. **Ari Leppälahti – Mikael Åkerblom,** Industrial Innovation in Finland. August 1991. 82 pp.

185. **Maarit Säynevirta,** Indeksiteoria ja ansiotasoindeksi. Lokakuu 1991. 95 s.

186. **Ari Tyrkkö,** Ahtaasti asuvat. Syyskuu 1991. 134 s.

187. **Tuomo Martikainen – Risto Yrjönen,** Voting, parties and social change in Finland. October 1991. 108 pp.

188. **Timo Kolu,** Työelämän laatu 1977–1990. Työn ja hyvinvoinnin koettuja muutoksia. Tammikuu 1992. 194 s.

189. **Anna-Maija Lehto,** Työelämän laatu ja tasa-arvo. Tammikuu 1992. 196 s.

190. **Tuovi Allén – Päivi Keinänen – Seppo Laaksonen – Seija Ilmakunnas,** Wage from Work and Gender. A Study on Wage Differentials in Finland in 1985. 88 pp.

191. **Kirsti Ahlqvist,** Kodinomistajaksi velalla. Maaliskuu 1992. 98 s.

192. **Matti Simpanen – Irja Blomqvist,** Aikuiskoulutukseen osallistuminen. Aikuiskoulutustutkimus 1990. Toukokuu 1992. 135 s.

193. **Leena M. Kirjavainen – Bistra Anachkova – Seppo Laaksonen – Iiris Niemi – Hannu Pääkkönen** – Zahari Staikov, Housework Time in Bulgaria and Finland. June 1992. 131 pp.

194. **Pekka Haapala – Seppo Kouvonen,** Kuntasektorin työvoimakustannukset. Kesäkuu 1992. 70 s.

195. **Pirkko Aulin-Ahmavaara,** The Productivity of a Nation. November 1992. 72 pp.

196. **Tuula Melkas,** Valtion ja markkinoiden tuolla puolen. Kanssaihmisten apu Suomessa 1980-luvun lopulla. Joulukuu 1992. 150 s.

197. **Fjalar Finnäs,** Formation of unions and families in Finnish cohorts born 1938–67. April 1993. 58 pp.

198. **Antti Siikanen – Ari Tyrkkö,** Koti – Talous – Asuntomarkkinat. Kesäkuu 1993. 167 s.

199. **Timo Matala,** Asumisen tuki ja aravavuokralaiset. Kesäkuu 1993. 84 s.

200. **Arja Kinnunen,** Kuluttajahintaindeksi 1990=100. Menetelmät ja käytäntö. Elokuu 1993. 89 s.

201. **Matti Simpanen,** Aikuiskoulutus ja työelämä. Aikuiskoulutustutkimus 1990. Syyskuu 1993. 150 s.

202. **Martti Puohiniemi,** Suomalaisten arvot ja tulevaisuus. Lokakuu 1993. 100 s.

203. **Juha Kivinen – Ari Mäkinen,** Suomen elintarvike- ja metallituoteteollisuuden rakenteen, kannattavuuden ja suhdannevaihteluiden yhteys; ekonometrinen analyysi vuosilta 1974 – 1990. Marraskuu 1993. 92 s.

204. **Juha Nurmela,** Kotitalouksien energian kokonaiskulutus 1990. Marraskuu 1993. 108 s.

205a. **Georg Luther,** Suomen tilastotoimen historia vuoteen 1970. Joulukuu 1993. 382 s.

205b. **Georg Luther,** Statistikens historia i Finland till 1970. December 1993. 380 s.

206. **Riitta Harala – Eva Hänninen-Salmelin – Kaisa Kauppinen-Toropainen – Päivi Keinänen – Tuulikki Petäjäniemi – Sinikka Vanhala,** Naiset huipulla. Huhtikuu 1994. 64 s.

207. **Wangqiu Song,** Hedoninen regressioanalyysi kuluttajahintaindeksissä. Huhtikuu 1994. 100 s.

208. **Anne Koponen,** Työolot ja ammattillinen aikuiskoulutus 1990. Toukokuu 1994. 118 s.

209. **Fjalar Finnäs,** Language Shifts and Migration. May 1994. 37 pp.

210. **Erkki Pahkinen – Veijo Ritola,** Suhdannekäänne ja taloudelliset aikasarjat. Kesäkuu 1994. 200 s.

211. **Riitta Harala – Eva Hänninen-Salmelin – Kaisa Kauppinen-Toropainen – Päivi Keinänen – Tuulikki Petäjäniemi – Sinikka Vanhala,** Women at the Top. July 1994. 66 pp.

212. **Olavi Lehtoranta,** Teollisuuden tuottavuuskehityksen mittaminen toimialatasolla. Tammikuu 1995. 73 s.

213. **Kristiina Manderbacka,** Terveydentilan mittarit. Syyskuu 1995. 121 s.

214. **Andres Vikat,** Perheellistyminen Virossa ja Suomessa. Joulukuu 1995. 52 s.

215. **Mika Maliranta,** Suomen tehdasteollisuuden tuottavuus. Helmikuu 1996. 189 s.

216. **Juha Nurmela,** Kotitaloudet ja energia vuonna 2015. Huhtikuu 1996. 285 s.

217. **Rauno Sairinen,** Suomalaiset ja ympäristöpolitiikka. Elokuu 1996. 179 s.

218. **Johanna Moisander,** Attitudes and Ecologically Responsible Consumption. August 1996. 159 pp.

219. **Seppo Laaksonen** (ed.), International Perspectives on Nonresponse. Proceedings of the Sixth International Workshop on Household Survey Nonresponse. December 1996. 240 pp.

220. **Jukka Hoffrén,** Metsien ekologisen laadun mittaaminen. Elokuu 1996. 79 s.

221. **Jarmo Rusanen – Arvo Naukkarinen** – Alfred Colpaert – Toivo Muilu, Differences in the Spatial Structure of the Population Between Finland and Sweden in 1995 – a GIS viewpoint. March 1997. 46 pp.

222. **Anna-Maija Lehto,** Työolot tutkimuskohteena. Marraskuu 1996. 289 s.

223. **Seppo Laaksonen** (ed.), The Evolution of Firms and Industries. June 1997. 505 pp.

224. **Jukka Hoffrén,** Finnish Forest Resource Accounting and Ecological Sustainability. June 1997. 132 pp.

225. **Eero Tanskanen,** Suomalaiset ja ympäristö kansainvälisestä näkökulmasta. Elokuu 1997. 153 s.

226. **Jukka Hoffrén,** Talous hyvinvoinnin ja ympäristöhaittojen tuottajana –Suomen ekotehokkuuden mittaaminen. Toukokuu 1999. 154 s.

227. **Sirpa Kolehmainen,** Naisten ja miesten työt. Työmarkkinoiden segregoituminen Suomessa 1970–1990. Lokakuu 1999. 321 s.

228. **Seppo Paananen,** Suomalaisuuden armoilla. Ulkomaalaisten työnhakijoiden luokittelu. Lokakuu 1999. 152 s.

229. **Jukka Hoffrén,** Measuring the Eco-efficiency of the Finnish Economy. October 1999. 80 pp.

230. **Anna-Maija Lehto – Noora Järnefelt** (toim.), Jaksaen ja joustaen. Artikkeleita työolotutkimuksesta. Joulukuu 2000. 264 s.

231. **Kari Djerf,** Properties of some estimators under unit nonresponse. January 2001. 76 pp.

232. **Ismo Teikari,** Poisson mixture sampling in controlling the distribution of response burden in longitudinal and cross section business surveys. March 2001. 120 pp.

233. **Jukka Hoffrén,** Measuring the Eco-efficiency of Welfare Generation in a National Economy. The Case of Finland. November 2001. 199 pp.

234. **Pia Pulkkinen,** "Vähän enemmän arvoinen" Tutkimus tasa-arvokokemuksista työpaikoilla. Tammikuu 2002. 154 s.

235. **Noora Järnefelt – Anna-Maija Lehto,** Työhulluja vai hulluja töitä? Tutkimus kiirekokemuksista työpaikoilla. Huhtikuu 2002. 130 s.

236. **Markku Heiskanen,** Väkivalta, pelko, turvattomuus. Surveytutkimusten näkökulmia suomalaisten turvallisuuteen. Huhtikuu 2002. 323 s.

237. **Tuula Melkas,** Sosiaalisesta muodosta toiseen. Suomalaisten yksityiselämän sosiaalisuuden tarkastelua vuosilta 1986 ja 1994. Huhtikuu 2003. 195 s.

238. **Rune Höglund – Markus Jäntti – Gunnar Rosenqvist (eds.),** Statistics, econometrics and society: Essays in honour of Leif Nordberg. April 2003. 260 pp.

239 **Johanna Laiho – Tarja Nieminen (toim.),** Terveys 2000 -tutkimus. Aikuisväestön haastatteluaineiston tilastollinen laatu. Otanta-asetelma, tiedonkeruu, vastauskato ja estimointi- ja analyysiasetelma. Maaliskuu 2004. 95 s.

Tutkimuksia-sarja kuvaa suomalaista yhteiskuntaa ja sen kansainvälistä asemaa tutkittujen tietojen pohjalta. Sarjassa julkaistaan Tilastokeskuksessa laadittuja tai Tilastokeskuksen aineistoihin perustuvia tieteellisiä tutkimuksia.

Terveys 2000 -tutkimus on Kansanterveyslaitoksen johdolla toteutettu suurtutkimus suomalaisten terveydentilasta. Sen tiedot kerättiin vuonna 2000 haastatteluiden, itsetäytettävien kyselyiden ja kliinisten tutkimusten avulla. Tässä raportissa kuvataan otanta-asetelmaa, haastatteluaineiston tiedonkeruuta ja Tilastokeskuksen tuottamien aineistojen laatua. Lisäksi tutkimuksessa verrataan eri analyysimenetelmiä ja perustellaan annetut suositukset Terveys 2000 -tutkimusaineiston tilastolliselle analysoinnille.

Tilastokeskus, myyntipalvelu
PL 4C
00022 TILASTOKESKUS
puh. (09) 1734 2011
faksi (09) 1734 2500
myynti@tilastokeskus.fi
www.tilastokeskus.fi

Statistikcentralen, försäljning
PB 4C
00022 STATISTIKCENTRALEN
tfn (09) 1734 2011
fax. (09) 1734 2500
myynti@stat.fi
www.stat.fi

Statistics Finland, Sales Services
P.O.Box 4C
FIN-00022 STATISTICS FINLAND
Tel. +358 9 1734 2011
Fax +358-9-1734 2500
myynti@stat.fi
www.stat.fi

ISSN 0355–2071
=Tutkimuksia
ISBN 952–467–267–7
Tuotenumero 89030
BE