HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta/Osasto — Fakultet/Sektion — Faculty	Laitos — Institution — Department
Matemaattis-luonnontieteellinen	Matematiikan ja tilastotieteen laitos

Tekijä — Författare — Author
Paula Bergman

Työn nimi — Arbetets titel — Title

Puuttuvuus GeneRISK-tutkimuksen perustietokyselyssä

Oppiaine — Läroämne — Subject
Tilastotiede

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Pro gradu -tutkielma	Syyskuu 2017	68 s. + liitteet 12 s.

Tiivistelmä — Referat — Abstract

Imputoinnilla tarkoitetaan sellaisia tilastollisia menetelmiä, joiden tarkoitus on täydentää puuttuvuutta sisältävää aineistoa. Puuttuvuus on iso ongelma tutkimuksissa, ja usein puuttuvat havainnot ja jopa kokonaiset havaintorivit jätetään huomiotta analyysejä tehdessä. Tämä voi kuitenkin merkittävästi vääristää analyysien tuloksia.

Tässä tutkielmassa esitellään erilaisia puuttuvuuden tyyppejä, käydään läpi puuttuvuuden mahdollisia syitä ja perehdytään erilaisiin imputointimenetelmiin. Imputointimenetelmien käyttöä havainnollistetaan esimerkeillä, jotka liittyvät GeneRISK-tutkimuksen perustietokyselyyn. GeneRISK-tutkimuksella pyritään selvittämään erityisesti sydän- ja verisuonitautien taustalla piileviä perinnöllisiä riskitekijöitä, sekä sitä, kuinka riskitiedon saaminen vaikuttaa yksilöiden myöhempään terveyskäyttäytymiseen.

Puuttuvuuden tyyppi vaikuttaa imputointimenetelmän valintaan, ja tutkielmassa esitellään kin niin täysin satunnainen, satunnainen, kuin ei-satunnainenkin puuttuvuus. Lisäksi sivutaan suunniteltua puuttuvuutta ja aineiston rakentamisvaiheessa syntyvää puuttuvuutta. Jos vastauksia puuttuu yksittäisiltä vastaajilta osasta kysymyksiä, on kyse erävastauskadosta, ja jos aineistosta puuttuu kokonaisia havaintorivejä, puhutaan yksikkövastauskadosta. Tutkielmassa keskitytään erävastauskatoon.

Tutkielmassa käytetään GeneRISK-tutkimuksen Kymenlaakson sairaanhoito- ja sosiaalipalvelujen kuntayhtymä Carean perustietokyselyaineistosta 18.1.2016 jäädytettyä otosta, ja sieltä valikoituja 1278 havaintoriviä. Tutkielmaan valikoitiin kiinnostuksen kohteena oleviksi muuttujiksi ruokailuun ja liikuntatottumuksiin liittyviä muuttujia, sekä taustamuuttujia. Aineistosta poistettiin kaikki sellaiset rivit, jotka sisälsivät puuttuvuutta näissä muuttujissa, ja tämän jälkeen täydelliseen aineistoon simuloitiin eri tyyppisiä puuttuvuuksia.

Puuttuvuutta pyrittiin korjaamaan niin yksinkertaisilla imputointimenetelmillä kuin kahdella erilaisella moni-imputointimenetelmälläkin. Yksinkertaisiin imputointimenetelmiin lukeutuu mm. mediaani-imputointi. Ehdollisten mallien moni-imputoinnin ja yhdistettyjen mallien moniimputoinnin on osoitettu kirjallisuudessa toimivan paremmin kuin yksinkertaisten imputointimallien, mutta tätä ei tässä tutkielmassa pystytty osoittamaan. Yhtenä syynä tähän saattaa olla kiinnostuksen kohteeksi valikoituneiden muuttujien väliset riippumattomuudet, sekä vastaajien keskinäinen samankaltaisuus. Edelleen hyvin yleinen tapa käsitellä puuttuvuutta on jättää se kokonaan huomiotta. Tutkielmassa kuitenkin huomataan, kuinka radikaaliin aineiston hupenemiseen se voi johtaa. Tutkielmassa osoitetaankin erityisesti se, kuinka tärkeää puuttuvuutta on tarkastella monelta eri kantilta aina puuttuvuuden syistä aineiston jatkokäyttötarkoituksiin asti.

Avainsanat — Nyckelord — Keywords
imputointi, moni-imputointi, puuttuvuus, erävastauskato, survey

Säilytyspaikka — Förvaringsställe — Where deposited
Kumpulan tiedekirjasto

Muita tietoja — Övriga uppgifter — Additional information

PUUTTUVUUS GENERISK-TUTKIMUKSEN PERUSTIETOKYSELYSSÄ

Paula Hannele Bergman

HELSINGIN YLIOPISTO
Matematiikan ja tilastotieteen laitos
Tilastotiede

Pro gradu -tutkielma

Syyskuu 2017

Sisältö

Sisältö **i**

1 Johdanto **1**

2 Puuttuvuus **3**

2.1 Puuttuvuuden tyypit 4

2.2 Muuttujien yhteydet vastaamattomuuden osalta 7

2.2.1 Käytännön esimerkki yhteyksien selvittämisestä 8

3 Imputointi **10**

3.1 Imputoinnin tavoitteet 11

3.2 Imputointimallin valinta 12

3.2.1 Malliluovuttaja (model-donor) ja vastaajaluovuttaja (real-donor) . 13

3.3 Yksinkertaisia imputointimalleja 14

3.3.1 Looginen imputointi 14

3.3.2 Keskilukuimputointi 14

3.4 Moni-imputointi 15

3.4.1 Ehdollinen moni-imputointi (FCS) 16

3.4.2 Yhdistettyjen mallien imputointi (JM) 19

3.4.3 Ehdollisen moni-imputoinnin ja JM-moni-imputoinnin vertailua . . 21

3.5 Muita mahdollisia menetelmiä 22

4 Esimerkki: GeneRISK-tutkimuksen kyselyaineisto **24**
4.1 Terveystietokyselyistä yleisesti . . . 25
4.2 Aineiston keruu . . . 26
4.3 Tutkimuksen vaiheet . . . 27
4.4 Aineiston muuttujat . . . 30
4.4.1 Väestöryhmään liittyvät muuttujat . . . 31
4.4.2 Elintapoihin liittyvät muuttujat . . . 32
4.5 Yksikkövastauskadon tarkastelu . . . 34
4.5.1 Osallistumiskriteerit . . . 34
4.5.2 Infopuhelimen kirjanpito . . . 36

5 Analyysit ja tulokset **38**
5.1 Erävastauskato tutkimusaineistossa . . . 38
5.1.1 Mistä puuttuvuus johtuu? . . . 43
5.2 Puuttuvuus kiinnostuksen kohteena olevissa muuttujissa . . . 44
5.3 Tutkimusaineiston imputointi . . . 46
5.3.1 Ehdollinen moni-imputointi . . . 48
5.3.2 JM-moni-imputointi . . . 49
5.3.3 Keskilukuimputointi . . . 49
5.3.4 CC-analyysi . . . 50
5.3.5 Looginen imputointi . . . 50
5.3.6 Tulokset . . . 51

6 Yhteenveto **63**

Viitteet **65**

Liitteet **69**

A Kysymyslomakkeen kysymysten aihepiirit **69**

B Kysymyskohtainen puuttuvuus **71**

C Tutkimuksen kulku **77**

D Kutsukirje **78**

Kiitokset

Kiitokset ohjaajilleni Kimmo Vehkalahdelle ja Samuli Ripatille, jotka ammattitaitoisen ja asiantuntevan ohjauksen lisäksi välittivät graduni edistymisestä ihan henkilökohtaisesti. Kiitos Matti Piriselle ja Sirkka-Liisa Varviolle, jotka olivat suureksi avuksi käytännön asioissa. Kiitos Johanna Arolle ja Elisabeth Widénille, jotka auttoivat minua lukuisissa GeneRISK-tutkimukseen liittyvissä kysymyksissä. Kiitos myös muille GeneRISK:in parissa työskenteleville ja työskennelleille, ilman heitä tätä gradua ei olisi koskaan syntynyt.

Kiitos kannustaville työkavereilleni. Kiitos Sannille, joka teki omaa graduaan aina yhden askelen minua edellä ja jolta sain korvaamatonta tukea. Kiitos Pauliinalle kullanarvoisista aamupalatreffeistä Gibbsin otantaan liittyen ja kaikista keskusteluistamme. Kiitos Heidille, jonka kanssa olemme saaneet yhdessä kokea graduprosessin ylä- ja alamäkiä. Kiitos Miikalle, joka luki ja kommentoi gradun raakiletta kaikkein ensimmäisenä, ja joka väsymättä auttoi kaikenlaisten siihen liittyvien ongelmien kanssa. Kiitos Iinalle, joka kannusti ja huolehti, etten väsy. Kiitos Moodin graduryhmälle, joka kokoontui vain pari kertaa, mutta josta sai mahtavaa vertaistukea. Kiitos perheelleni, joka uskoi, että pystyn tähän.

Kiitos kaikille ystäville, tutuille ja tutuntutuille, jotka ovat olleet kiinnostuneita gradustani myös silloin, kun oma motivaationi meinasi hiipua. Kiitos kaikille, jotka ovat olleet elämässäni tämän prosessin aikana.

Luku 1

Johdanto

Puuttuvuus on tutkimuksessa varsin yleistä ja johtaa usein ongelmiin tiedon analysointivaiheessa. Puuttuvuudella tarkoitetaan sitä, että aineistosta puuttuu kokonaisia havaintorivejä tai niiden osia. Puuttuvuuden syntymiseen johtavat lukuisat seikat ja niiden tunteminen auttaa valitsemaan parhaat kadonhallintamenetelmät.

Tässä tutkielmassa käsitellään puuttuvuutta kyselytutkimuksessa. Aihetta lähestytään GeneRISK-tutkimuksen perustietokyselyn avulla. Käytössä on tutkimuksen Kymenlaakson sairaanhoito- ja sosiaalipalvelujen kuntayhtymän, Carean, aineisto (N=1400, tilanne 18.1.2016). Tästä poistetaan vielä rivit, joilta puuttuu havaintoarvoja myöhemmin esiteltäviin taustamuuttujiin. Tämän jälkeen saatu uusi N=1278. Kyselylomakkeista tarkastellaan sitä, mihin kysymyksiin on jätetty vastaamatta, mistä tämä mahdollisesti johtuu ja mitä asialle voidaan tehdä. Kiinnostuksen kohteena ovat erityisesti liikunta- ja ruokailutottumuksia koskevat muuttujat.

Puuttuvuutta on montaa eri tyyppiä, ja siksi onkin tärkeää tarkastella käytettävissä olevan tiedon valossa puuttuvuutta kysymyskohtaisesti. Puuttuvuuden tyyppien tarkastelun lisäksi tutkielmassa pohditaan sitä, kuinka dataa voitaisiin tilastollisesti täydentää siten, että puuttuvat arvot muuttujissa eivät vääristäisi datan analysointia. Puuttuvien tietojen käsittelymenetelmiin vaikuttavat eritoten puuttuvuuden tyyppi, käytettävissä olevat

taustatiedot, sekä muihin kysymyksiin vastaaminen tai muiden osallistujien vastaukset.

Puuttuvuus-luvussa perehdytään puuttuvuuteen käsitteenä, sekä esitetään määritelmiä puuttuvuuden tyypeille. Luvun lopuksi esitellään esimerkin omaisesti yksi tapa selvittää puuttuvuuden tyyppiä. Seuraavaksi esitellään yleisesti käytettäviä menetelmiä puuttuvuutta sisältävän datan käsittelemiseen. Tämän jälkeen tarkastellaan, millaista puuttuvuutta tutkittava data sisältää ja esittellään esimerkkien avulla käsittelymenetelmiä ja niiden käyttöä oikeassa tilanteessa.

Puuttuvuutta on myös se, jos joku on jättänyt kokonaan osallistumatta tutkimukseen. Niinpä kiinnostuksen kohteena on selvittää myös sitä, kuinka moni tutkimukseen kutsutuista on osallistunut tutkimukseen, sekä syitä sille, miksi tutkimukseen osallistumattomat ovat mahdollisesti päättäneet jättää osallistumatta. Tämän selvittämiseksi on tarkasteltu Carean infopuhelinaineistoa, josta käy ilmi yllä mainittuja seikkoja.

Aineiston puuttuvuuden kartoittamisen jälkeen esitellään menetelmiä puutteellisen datan laadun parantamiseksi ja esitetään esimerkkejä siitä, kuinka tutkimusdataa voidaan käsitellä siten, että jatkoanalyysit ovat mahdollisia. Puuttuvuuden käsittelyssä keskeistä on tietää, millaiseen tarkoitukseen dataa käytetään käsittelyn jälkeen, jotta aineistoa voidaan täydentää sen mukaisilla menetelmillä.

Esimerkiksi Pentala (2014) on käsitellyt pro gradu -tutkielmassaan tilastollisia kadonhallintamenetelmiä. Hänen mukaansa nykyään katoa esiintyy yhä enenevässä määrin, ja puuttuvat tiedot jätetään usein huomiotta analyyseissä, mikä johtaa siihen, ettei tutkimus enää edusta sitä väestöä, jota sen oli alunperin tarkoitus edustaa. Tässä tutkielmassa on tarkoitus edistää niiden menetelmien käyttöä, jotka säilyttäisivät aineiston laadultaan mahdollisimman korkeatasoisena, jotta se vastaisi tutkimuksen tarpeisiin.

Tutkielman lopussa pohditaankin saatuja tuloksia ja menetelmien hyödyllisyyttä tässä aineistossa, sekä päätelmien yleistettävyyttä.

Luku 2

Puuttuvuus

Puuttuvuudella tarkoitetaan kaikenlaisia aineistossa olevia puutteita, sekä niitä puutteita, jotka saadaan tavoitellun aineiston ja saavutetun aineiston erotuksena. Puuttuvuus voi aiheuttaa monenlaisia ongelmia aineiston käytettävyyteen ja pahimmillaan estää luotettavien analyysien tekemisen kokonaan. Puuttuvuuden syiden selvittäminen on tärkeää, jotta voidaan löytää oikea lähestymistapa tilanteen korjaamiseksi. Puuttuvuutta on monenlaista, eikä kaikki puuttuvuus ole aina sellaista, josta olisi pyrittävä eroon.

Määritellään seuraavaksi muutamia oleellisesti puuttuvuuteen liittyviä käsitteitä.

Määritelmä 2.1. ***Yksikkövastauskato***
Yksikkövastauskato (unit non-response) tarkoittaa puuttuvuutta silloin, kun tutkittavalta ei ole saatu lainkaan vastauksia mihinkään kysymykseen. Tällöinkin vastaajasta saattaa kuitenkin löytyä kehikkotasoista tietoa, kuten syntymäaika. Myös silloin, kun täytetty lomake on lukukelvoton tai muuten tarpeeksi puutteelliseksi todettu, voidaan puhua yksikkövastauskadosta. Tästä käytetään myös nimeä täydellinen vastaamattomuus. [Laaksonen, 2013]

Tässä tutkielmassa käytettävässä tutkimusaineistossa ei varsinaista yksikkövastauskatoa esiinny, koska tutkimushenkilöiden rekrytointia jatketaan, kunnes haluttu määrä saavutetaan; jos kutsutut eivät osallistu, kutsutaan lisää ihmisiä. Kaikki kutsutut eivät ole

kuitenkaan lähteneet mukaan tutkimukseen, joten siinä mielessä yksikkövastauskadosta voidaan puhua.

Määritelmä 2.2. ***Erävastauskato***
Erävastauskadolla (item non-response) tarkoitetaan tilannetta, jossa puuttuvuutta on yksittäisen muuttujan kohdalla. Tätä voidaan kutsua myös *osittaiseksi vastaamattomuudeksi*, eli vastauksia puuttuu yksittäisiltä vastaajilta osittain. Joskus erävastauskato voi olla myös seurausta esimerkiksi epäselvistä tai vaikeaselkoisista vastauksista, jolloin tällaiset vastaukset joudutaan käsittelemään puuttuvina tietoina. [Laaksonen, 2013]

Määritelmä 2.3. ***Suunniteltu puuttuvuus***
Suunniteltu puuttuvuus viittaa tässä tutkielmassa sellaiseen puuttuvuuteen, jossa tutkija on tietoisesti jättänyt keräämättä vastauksia joihinkin tiettyihin kysymyksiin tiettyjen vastaajien kohdalla. Tällaisesta on kyse esimerkiksi silloin, kun miehiltä ei kerätä vastauksia naisia koskeviin kysymyksiin tai kun tupakoimattomilta ei kerätä tietoa heidän tupakointitottumuksistaan. Tällaisia kysymyksiä edeltää yleensä jokin kartoittava kysymys, joka jakaa vastaajat sen mukaan, tarvitseeko heidän vastata jatkokysymyksiin vai ei.

2.1 Puuttuvuuden tyypit

Puuttuvia arvoja voidaan käsitellä monin tavoin. Käsittelyyn vaikuttaa erityisesti se, onko puuttuvuus satunnaista vai ei-satunnaista, ja mihin aineistoa halutaan käyttää.

Määritellään seuraavaksi teoreettinen kehys puuttuvuuden tyypeille. Määrittely perustuu teoksen [Little and Rubin, 2002] lukuun 1.3.

Merkitään täydellistä aineistoa $Y = (y_{ij})$ ja puuttuvan aineiston indikaattorimatriisia $M = (M_{ij})$. Y on tässä i x j -matriisi, jossa rivit vastaavat havaintoja ja sarakkeet muuttujia. Indikaattorimatriisi on samansuuruinen kuin datamatriisi, ja matriisin solu saa arvon 1, jos kyseinen muuttuja on havaittu ja 0, jos arvo puuttuu. Se siis nimensä mukaisesti osoittaa (indicate) puuttuvuuden aineistosta. Otetaan esimerkiksi kuviteltu

aineisto, jossa on neljä muuttujaa ja viisi havaintoa. Taulukossa 2.1 nähdään alkuperäinen datamatriisi ja taulukossa 2.2 siitä muodostettu puuttuvuuden indikaattorimatriisi. Muuttujia merkitään m1-m4.

Taulukko 2.1: Alkuperäinen datamatriisi.

havainto \muuttuja	m1	m2	m3	m4
havainto1	13	NA	1	0
havainto2	18	NA	0	5
havainto3	15	2	3	NA
havainto4	18	1	3	4

Taulukko 2.2: Puuttuvuuden indikaattorimatriisi.

havainto \muuttuja	m1	m2	m3	m4
havainto1	1	0	1	1
havainto2	1	0	1	1
havainto3	1	1	1	0
havainto4	1	1	1	1

Aineistossa on siis kolme puuttuvaa arvoa, kaksi muuttujassa 2 ja yksi muuttujassa 4.

Merkitään puuttuvuuden mekanismia nyt M:n jakaumana ehdolla Y, jonka tiheysfunktiota voidaan merkitä $f(M \mid Y, \phi)$, missä ϕ kuvaa tuntemattomia parametreja.

Määritelmä 2.4. ***Täydellisesti satunnainen puuttuvuus*** **(MCAR)**
Puuttuvuutta sanotaan täydellisesti satunnaiseksi (missing completely at random), kun se ei riipu datan Y arvoista lainkaan, eli

$$f(M \mid Y, \phi) = f(M \mid \phi) \text{ kaikilla } Y, \phi.$$

Yhden muuttujan Y tapauksessa voitaisiin muodostaa yksittäinen indikaattorimuuttuja M, joka saa arvon 1, kun Y:n arvo puuttuu ja 0 kun arvo on havaittu. Jos puuttuvuuden todennäköisyys on sama kaikilla y_i eli

$$P(M_i = 1 \mid y_{i1}, ..., y_{iK}; \phi) = \phi,\ i \in \{1, \ldots, N\}$$

niin tällöin havaitut arvot ovat satunnaisotos kaikista arvoista.

Esimerkki 2.5. Kuvitellaan kolmen muuttujan aineisto, jossa kiinnostuksen kohteena olevana muuttujana on koulutustaso, muiden muuttujien ollessa ikä ja sukupuoli. Jos koulutustason puuttumisen todennäköisyys ei riipu mistään näistä kolmesta muuttujasta, on puuttuvuuden tyyppi tällöin MCAR.

Määritelmä 2.6. ***Satunnainen puuttuvuus*** **(MAR)**
Merkitään Y_{obs} havaittuja Y:n arvoja ja Y_{mis} puuttuvia Y:n arvoja. Satunnaisesta puuttuvuudesta on kysymys silloin, kun puuttuvuus riippuu vain havaituista Y:n arvoista, eli

$$f(M \mid Y, \phi) = f(M \mid Y_{obs}, \phi) \text{ kaikilla } Y_{mis}, \phi,$$

eli tiheysfunktio on sama, otettiin puuttuvat arvot mukaan tai ei.

Kuvitellaan esimerkin 2.5 aineiston tilanne. Jos nyt koulutustason puuttuminen riippuu iästä ja sukupuolesta, mutta ei koulutustasosta, on puuttuvuuden tyyppi MAR.

Määritelmä 2.7. ***Ei-satunnainen puuttuvuus*** **(NMAR)**
Jos M:n jakauma riippuu myös puuttuvista Y:n arvoista, on kyse ei-satunnaisesta puuttuvuudesta.

Kuvitellaan jälleen esimerkin 2.5 aineiston tilanne. Nyt jos todennäköisyys sille, että koulutustaso on havaittu, riippuu koulutustasosta samassa ikä- ja sukupuoliryhmässä, on kyse ei-satunnaisesta puuttuvuudesta NMAR.

Aineiston käsittelyn kannalta ideaalein tilanne olisi, että puuttuvuus olisi täysin satunnaista, eli määritelmän 2.4 mukaista. Kuitenkin useimmiten joudutaan tyytymään määritelmän 2.6 mukaiseen oletukseen. Oletusta on käytännössä hankala vahvistaa, mutta esimerkiksi sisällyttämällä analyysiin useita täydellisesti havaittuja, epätäydellistä muuttujaa selittäviä muuttujia, voidaan vähentää muuttujan riippuvuutta sen puuttuvista arvoista. [Gelman et al., 1995]

2.2 Muuttujien yhteydet vastaamattomuuden osalta

Kysymyksiin vastaamattomuus voi olla systemaattista: Tutkimukseen osallistuva henkilö voi esimerkiksi vastata vain jokaisen kysymyssarjan pakollisiin kysymyksiin ja jättää muihin vastaamatta. Osallistuja voi väsyä lomakkeen täyttämiseen ja jättää loppua kohden enemmän vastaamatta kuin kysymyslomakkeen alkupäässä. Lisäksi henkilö saattaa jättää tarkoituksella vastaamatta esimerkiksi osaan alkoholi- ja tupakointikysymyksistä. Tähän syynä voi olla vaikkapa häpeä omista elintavoista, tai jopa kiinnijäämisen pelko, esimerkiksi huumeiden käyttöön liittyvien kysymysten osalta.

Puuttuvuus ei ole tällaisissa tilanteissa satunnaista, eli kysymys on määritelmän 2.7 mukaisesta, ei-satunnaisesta, puuttuvuudesta. Tällöin tavallisimpia puuttuvan datan käsittelymenetelmiä ei voida hyödyntää. Tämän vuoksi onkin tärkeää pyrkiä selvittämään, onko puuttuvuus ei-satunnaista. Asian selvittämiseksi voidaan muun muassa tarkastella eri kysymyksiin vastaamattomuuden yhteyksiä: Onko esimerkiksi niin, että ne, jotka eivät vastaa liikuntatottumuskysymyksiin, eivät vastaa myöskään ruokailutottumuskysymyksiin tai ne, jotka eivät vastaa alkoholinkäyttökysymyksiin, eivät myöskään vastaa tupakointikysymyksiin?

Kiinnostuksen kohteena tällaisissa pohdinnoissa on se, korreloiko tiettyihin kysymyksiin vastaamattomuus muiden kysymysten vastaamattomuuden kanssa. Tätä voidaan selvittää esimerkiksi luomalla niin kutsuttuja dummy-muuttujia. Dummy-muuttuja on dikotominen, eli kaksiarvoinen, apumuuttuja, joka voisi saada arvon 0, kun henkilö on jättänyt

vastaamatta kysymykseen, ja arvon 1, kun henkilö on vastannut kysymykseen. Jos tarkastelu tehtäisiin kaikille datan muuttujille, olisi kyse samankaltaisen indikaattorimatriisin rakentamisesta, kuin mitä on esitelty luvussa 2.1.

Kahden muuttujan puuttuvuuden välistä yhteyttä voidaan tarkastella dummy-muuttujien avulla esimerkiksi χ^2-testillä. χ^2-testi on testausmenetelmä kategoristen muuttujien välillä vallitsevan yhteyden selvittämiseksi. Usein testiä käytetään nimenomaan kaksiarvoisille muuttujille. Testisuure lasketaan havaittujen frekvenssien ja odotettujen frekvenssien avulla. Mitä pienemmän p-arvon testisuure saa, sitä enemmän testi antaa tukea sille, että muuttujat riippuvat toisistaan. [Taanila, 2011]

2.2.1 Käytännön esimerkki yhteyksien selvittämisestä

Muodostettiin indikaattorimatriisi ”puute”, joka on samankokoinen kuin tutkittava aineisto, eli siinä on yhtä monta riviä ja yhtä monta saraketta. Indikaattorimatriisin jokainen sarake kuvastaa tutkimusaineiston muuttujaa ja jokainen rivi havaintoa. Jos kyseisen muuttujan arvoa ei ole havaittu, solu saa arvon 1 ja muutoin 0.

Tämän jälkeen testattiin χ^2-testin avulla, onko sarakkeiden arvojen välillä yhteyksiä, eli onko muuttujien sisältämän puuttuvuuden välillä yhteyksiä.

Selvitettiin, onko tutkimusaineiston muuttujien

- 5.4B Kuinka monta tuntia istutte keskimäärin arkipäivänä? kotona televisiota tai videoita katsellen

ja

- 11.3G Kuinka usein tavallisesti käytätte seuraavia elintarvikkeita? Tuoresalaattia, tuoreita kasviksia

puuttuvuuden välillä yhteyttä.

Taulukko 2.3: Vastaamatta jättämisen yhteyksiä. Havaitut frekvenssit

5.4B \11.3G	ei puutu	puuttuu
ei puutu	1143	10
puuttuu	122	3

Taulukko 2.4: Vastaamatta jättämisen yhteyksiä. Odotetut frekvenssit

5.4B \11.3G	ei puutu	puuttuu
ei puutu	1141.3	11.7
puuttuu	123.7	1.3

χ^2-testi antaa p-arvoksi 0.249. Suurehko p-arvo antaa tukea väitteelle, että muuttujien puuttuvuuksien välillä ei olisi yhteyttä. Molempiin kysymyksiin vastaamattomien henkilöiden lukumäärä on toisaalta hyvin pieni, mikä luo epävarmuutta testin p-arvon tulkintaan. Fisherin testi, jota pidetään yleisesti suositeltavampana, vastaavaan tarkoitukseen sopivana testinä, kun jonkin solun frekvenssi on pienempi kuin viisi, antaa p-arvoksi 0.126. Tätä ei myöskään voida pitää tilastollisesti merkitsevänä.

Luku 3

Imputointi

Seuraava luku perustuu teokseen [Laaksonen, 2013], ellei toisin ole mainittu. Imputoinnilla tarkoitetaan aineiston puuttuvien tai käyttökelvottomien arvojen täydentämistä tilastollisin menetelmin. Yksi tavallisimmista imputoinnin taustalla olevista syistä on puuttuneisuus tietyn vastaajan kohdalla tietyissä kysymyksissä, eli erävastauskato (ks. määritelmä 2.2). Imputointiin voi johtaa myös se, että mitattu arvo vaikuttaa selkeästi virheelliseltä. Yksikkövastauskatoakin (ks. määritelmä 2.1) voidaan imputoida *massaimputointia* hyödyntäen. Massaimputoinnilla tarkoitetaan suurien, useita muuttujia käsittävien puuttuvuusrivien tilastollista täydentämistä. Tämä on tyypillistä esimerkiksi yritystutkimuksissa. Imputointimenetelmiä voidaan hyödyntää myös sellaisissa tilanteissa, joissa havaintoyksikkö olisi pääteltävissä saadun arvon perusteella. Tällöin imputointia voidaan käyttää aineiston suojaamiseksi ja tietosuojan takaamiseksi. Tässä tutkielmassa käsitellään ainoastaan ensimmäisenä mainittua tilannetta, eli tavallisen erävastauskadon tilastollista täydentämistä.

Jos aineistoa kerättäessä on saatu tietoa myös vastaamattomuuden syistä, voidaan käyttää apuna niin sanottua imputointikoodausta. Tällöin puuttuvan arvon tilalle voidaan antaa arvo, joka kuvastaa syytä, jonka vuoksi tieto puuttuu. Syynä voisi olla esimerkiksi se, ettei tutkittava osannut vastata, ettei tutkittava halunnut vastata tai se, ettei kysymys koskenut vastaajaa. Näitä tietoja voidaan hyödyntää myös imputoinnissa, koska vastaa-

mattomuuden syyt voivat jakaa tutkittavia ryhmiin, joille voidaan käyttää eri imputointimenetelmiä. Tutkielmassa käytettävässä aineistossa tämä ei kuitenkaan kiinnostuksen kohteena olevien muuttujien osalta ole tiedossa.

Määritelmä 3.1. ***Deterministinen imputointi***
Deterministisestä imputoinnista on kysymys silloin kun imputoinnilla päädytään aina samaan lopputulokseen. Tällöin ei voida suorittaa moninkertaista imputointia.

Määritelmä 3.2. ***Stokastinen imputointi***
Stokastinen imputointi tarkoittaa satunnaislukujen avulla tuotettua imputointia. Eri satunnaisluvuilla imputointi tuottaa eri arvon, mikä mahdollistaa myös moninkertaisen imputoinnin ja sitä kautta epävarmuuden huomioimisen.

3.1 Imputoinnin tavoitteet

Imputoinnilla on oltava selkeä tavoite, kuten esimerkiksi estimoinnin parantaminen. Imputoinnin tarpeellisuutta mietittäessä on otettava huomioon, että sen seurauksena aineiston muut estimaatit saattavat huonontua. Tilanteesta riippuen saattaa siis joskus olla viisainta jättää imputointi tekemättä. Imputoinnin tavoitteena ei useinkaan ole muodostaa parhaita arvoja yksittäisten arvojen korvaamiseksi, vaan paikata puuttuvia arvoja sellaisilla arvoilla, joiden avulla muuttujia voidaan hyödyntää parametrien päättelyyn. [Andridge, 2010]

Imputointiin vaikuttaa myös se, millaiseen analyysiin tietoja halutaan hyödyntää, eli millaisiin tutkimuskysymyksiin pyritään vastaamaan. Jos esimerkiksi kysyttäisiin vastaajien alkoholinkäyttötottumuksia, voitaisiin ensin kysyä ”Käytätkö alkoholia?”, jonka jälkeen seuraavat kysymykset osoitetaan vain myöntävästi vastanneille. Jos seuraava kysymys on ”Kuinka monta alkoholiannosta juot viikossa seuraavia juomia?” ja alakohdat esimerkiksi ”miedot alkoholijuomat”, ”viini”, ”väkevät alkoholijuomat”, voidaan alkoholia käyttämättömien kohdalla arvo imputoida tarpeen mukaan esimerkiksi arvolla 0 tai NA (not available = ei saatavilla). Jos tutkimuskysymyksenä olisi vaikkapa ”Kuinka paljon alkoholia vastaajat käyttävät keskimäärin viikossa?”, on tarpeen käyttää alkoholia juomattomien kohdalla

arvoa 0. Jos taas ollaan kiinnostuneita vain siitä, mikä on esimerkiksi eri alkoholijuomien määrien suhde, on syytä käyttää vain edelliseen kysymykseen myöntävästi vastanneita.

Parhaimmassa tapauksessa imputoitu aineisto olisi arvoiltaan mahdollisimman lähellä todellista aineistoa. Harvoin tämä on kuitenkaan mahdollista. Usein hyvyyden kriteerinä voidaankin käyttää sitä, että imputoidun aineiston jakauma olisi mahdollisimman lähellä aineiston todellista jakaumaa. Todellisissa puuttuvan aineiston tilanteissa on useimmiten mahdotonta tietää, mikä oikea jakauma olisi, ja siksi joudutaankin tyytymään jakauman uskottavuuden tarkasteluun. Myös muuttujien välisten yhteyksien tarkastelu on tärkeää, erityisesti jos aineistoa toivotaan jatkossa voitavan analysoida monimuuttujamenetelmien avulla.

3.2 Imputointimallin valinta

Aluksi määritellään muuttujat, joita halutaan käsitellä. Näistä muuttujista valitaan ne arvot, joita pyritään imputoimaan. Tarpeelliset taustamuuttujat ja apumuuttujat täytyy käydä läpi, jotta niitä voidaan hyödyntää imputoinnissa. Tämän jälkeen voidaan määritellä imputointimalli.

Mallin voi muodostaa joko harkinnan perusteella tai estimoimalla. Malli voi olla joko deterministinen tai stokastinen, deterministisen ollessa kuitenkin yleisempi. Mallia muodostettaessa määritellään yhteys apumuuttujien x ja imputoitavan muuttujan y välille. Estimointi voidaan perustaa aineistoon tai aiempaan vastaavaan aineistoon. Mallia estimoitaessa ihannetilanne olisi, ettei selittävissä muuttujissa olisi puuttuvuutta. Selitettäväksi muuttujaksi voidaan valita joko ”muuttuja, jonka arvoja imputoidaan” tai ”puuttuneisuuden kaksiarvoinen indikaattori muuttujalle, jota imputoidaan.” Selittävät muuttujat voivat olla kategorisia tai jatkuvia, mutta jatkuvien muuttujien tapauksessa selkeästi poikkeavia havaintoja ei saisi olla. Imputoinnissa voidaan hyödyntää myös homogeenisia osajoukkoja, eli imputointisoluja.

3.2.1 Malliluovuttaja (model-donor) ja vastaajaluovuttaja (real-donor)

Varsinaisessa imputointivaiheessa voidaan valita joko malliluovuttaja-, tai vastaajaluovuttajamenetelmä. Malliluovuttajamenetelmässä imputoidut arvot muodostetaan nimensä mukaisesti mallista saatavien arvojen perusteella, joko deterministisesti tai stokastisesti. Vastaajaluovuttajamenetelmässä imputoitu arvo saadaan joltain toiselta vastaajalta. Lähestymistapa on valittava tilannekohtaisesti riippuen myös muuttujan luonteesta. Esimerkiksi vastaajan tuloja imputoitaessa malliluovuttajamenetelmä saattaisi tuottaa negatiivisia arvoja, mikä ei olisi hyväksyttävää. Tämä täytyy ottaa huomioon mallia muodostettaessa.

Vastaajaluovuttajan kohdalla täytyy harkita, miten luovuttaja valitaan. Joskus voidaan olettaa, että kaikki vastaajat joukossa tai imputointisolussa ovat keskenään yhtä mahdollisia, jolloin arvon luovuttaja voitaisiin valita satunnaisesti. Imputointisolulla tarkoitetaan tässä aineiston havaintoja, joiden joukosta arvo valitaan. Tällöin on kyse ns. hot deck -imputoinnista. Imputointisolussa on oltava puuttuvien arvojen ohella myös oikeita havaintoarvoja, jotta imputointi olisi mahdollista. Läheisin arvo voidaan myös valita aputiedon avulla tai muodostamalla euklidinen etäisyysmitta muuttujista. Läheisimmän arvon voi myös hakea malliluovuttajamenetelmän arvojen avulla. Jälkimmäisessä tilanteessa on siten tehtävä sekä malliluovuttaja-, että vastaajaluovuttajamenetelmää hyödyntävä imputointi. Mikäli malliluovuttajamenetelmällä löydetään sopiva arvo, ei vastaajaluovuttajamenetelmään tarvitse jatkaa.

3.3 Yksinkertaisia imputointimalleja

Seuraavassa perehdytään muutamiin erilaisiin, yleisesti käytettäviin, imputointimalleihin.

3.3.1 Looginen imputointi

Puuttuvia arvoja voidaan yksinkertaisimmillaan imputoida loogisen päättelyn avulla: Jos puuttuvan muuttujan arvo pystytään päättelemään jonkin muun muuttujan arvon perusteella, voidaan tätä tietoa hyödyntää puuttuvan arvon täydentämiseksi. Monissa kyselylomakkeissa esimerkiksi esiintyy kysymys sekä iästä, että syntymäajasta. Jos ikää koskevaan kysymykseen on jätetty vastaamatta, voidaan arvo laskea syntymäajan perusteella. Toisena esimerkkinä voidaan ajatella, että henkilö on jättänyt vastaamatta kysymykseen siitä, onko hänellä koskaan todettu raskausajan diabetesta ja toisaalta vastannut kieltävästi kysymykseen siitä, onko koskaan ollut raskaana. Tästä voidaan tehdä päätelmä siitä, että raskausajan diabetesta ei ole todettu. [Gelman and Hill, 2007]

3.3.2 Keskilukuimputointi

Keskilukuimputoinnissa puuttuvien arvojen paikalle sijoitetaan aineistosta laskettu muuttujan keskiarvo tai mediaani. Se on siten deterministinen menetelmä. Keskilukuimputointia ei yleisesti pidetä kovinkaan hyvänä menetelmänä, koska se saattaa merkittävästi vääristää muuttujan jakaumaa ja sitä kautta vääristää myös keskeisiä tunnuslukuja ja aliarvioida keskihajontaa. Yksi imputoinnin keskeisiä tavoitteita on myös säilyttää muuttujien keskinäiset suhteet todenmukaisina ja keskilukuimputoinnin seurauksena tämä saattaa vaikeutua. [Gelman and Hill, 2007]

Lisäksi keskiluvun laskeminen ei ole mielekästä karakterisille muuttujille, vaikka niitä olisikin aineistossa merkitty numeroarvoilla. Esimerkiksi jos aineistossa merkitään värejä sininen = 1, keltainen = 2, punainen = 3, ei niiden keskiarvo luonnollisestikaan ole keltainen.

3.4 Moni-imputointi

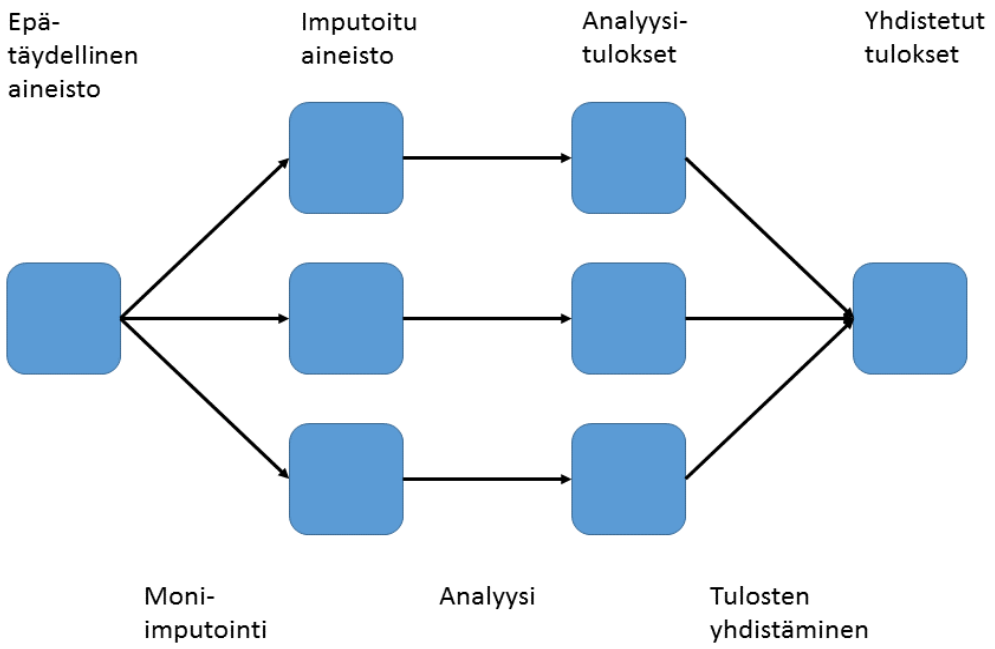
Moni-imputoinnilla tarkoitetaan useampien imputointikertojen tekemistä puuttuville arvoille siten, että saavutetaan useampia täydellisiä datoja. Moni-imputoinnin etuna tavalliseen yksinkertaiseen imputointiin verrattuna on se, että parametriestimaatit eivät sisällä niin paljon vaihtelua. Kun otetaan keskiarvo useampien imputoitujen aineistojen yli lasketuista estimaateista, vaihtelun määrä vähenee huomattavasti. Lisäksi useiden aineistojen yli lasketut estimaattien vaihtelut tarjoavat mahdollisuuden laskea puuttuvien arvojen epävarmuutta kuvaavia keskivirheitä. [Allison, 2009]

Moni-imputointi ei ole siinä mielessä oma imputointimenetelmänsä vaan ennemminkin tapa suorittaa imputointia. Sen sijaan, että imputoitaisiin vain kerran ja saavutettaisiin yksi täydellinen data, imputointi tehdään useita kertoja, jotta otettaisiin huomioon imputoinnin mukanaan tuoma vaihtelu.

Moni-imputoinnissa on kolme päävaihetta:

1. imputointi
2. analyysi
3. yhdistäminen

Aluksi havaitaan epätäydellinen aineisto, joka sisältää puuttuvia arvoja. Tämän jälkeen imputoidaan haluttu määrä täydellisiä aineistoja tästä epätäydellisestä aineistoista ja sen jälkeen suoritetaan analyysit kaikille imputoiduille aineistoille. Lopuksi tuloksena saadut estimaatit yhdistetään yhdeksi estimaatiksi laskemalla niistä keskiarvo, ja tälle lasketaan varianssi. Vaiheet on esitelty kuvassa 3.1. [van Buuren and Groothuis-Oudshoorn, 2011] Moni-imputoinnissa on kaksi yleistä etenemistapaa: Ehdollinen moni-imputointi (FCS) ja yhdistettyjen mallien moni-imputointi (JM-moni-imputointi). Näitä esitellään seuraavaksi.



Kuva 3.1: Moni-imputoinnin vaiheet. Kuva on laadittu [van Buuren and Groothuis-Oudshoorn, 2011] pohjalta.

3.4.1 Ehdollinen moni-imputointi (FCS)

Ehdollinen moni-imputointi (fully conditional specification) on iteratiivinen menetelmä, joka suoritetaan muuttuja kerrallaan ehdollisten jakaumien avulla. Menetelmä sopii kaikenlaisille muuttujille, myös kategorisille. Imputointiin käytetään myös ketjutettujen yhtälöiden moni-imputointialgoritmia (multiple imputation of chained equations, MICE).

Koska mallin muuttujat ovat keskenään erilaisia, on hyvä määrittää imputointimalli erikseen jokaiselle muuttujalle. Esimerkiksi jatkuvia muuttujia voidaan imputoida eri tavoin kuin diskreettejä. Tähän tarpeeseen on kehitetty ketjutettujen yhtälöiden menetelmä, jos-

sa edetään muuttuja kerrallaan.

Ehdollisessa moni-imputoinnissa jokaiselle osittain havaitulle muuttujalle muodostetaan ehdollinen jakauma iteratiivisesti mallintamalle, ehdollistaen sekä havaituilla, että jo imputoiduilla muiden muuttujien arvoilla. Malli korostaa sitä, että suuri osa tilastollisista malleista on nimenomaan ehdollisia malleja.

Algoritmi koostuu kahdesta silmukasta, joista sisempi imputoi muuttujia ennustamalla niiden arvoja muiden mallin muuttujien avulla. Mallin muita muuttujia käytetään siis täydellisinä ennustemuuttujina imputoitavalle muuttujalle. Jos puuttuvuutta sisältäviä muuttujia merkitään $Y_1, .., Y_p$ ja täydellisesti havaittuja selittäviä muuttujia X, niin aluksi imputoidaan Y_1 ehdolla $Y_2, .., Y_p$ ja X, sitten imputoidaan Y_2 ehdolla Y_1,$Y_3, .., Y_p$ ja X ja niin edelleen. [Gelman and Hill, 2007]. Ulompi silmukka toistaa tätä prosessia, kunnes ehdolliset jakaumat vaikuttavat supenneen siten, että imputointiarvoja poimitaan samasta yhteisjakaumasta. [Kropko et al., 2014] Kyseessä on bayesiläinen lähestymistapa, jossa priorijakaumasta muodostuu lopulta posteriorijakauma, kun mallia täydennetään jatkuvasti iteraatioista saatavilla uusilla tiedoilla.

Oletetaan, että Y_j, missä $(j = 1, ..., p)$, on yksi p:stä puuttuvuutta sisältävästä muuttujasta ja $Y = (Y_1, ..., Y_p)$. Merkitään havaittuja muuttujien arvoja Y_j^{obs} ja ei-havaittuja Y_j^{mis}. Imputointien lukumäärää voidaan merkitä $m \geq 1$ ja h:nnetta imputoitua aineistoa $Y^{(h)}$, missä $h = 1, .., m$. $p - 1$ Y-muuttujaa Y_j:tä lukuunottamatta merkitään $Y_{-j} = (Y_1, ..., Y_{j-1}, Y_{j+1}, ..., Y_p)$. Olkoon Q kiinnostuksen kohteena oleva arvo (kuten esimerkiksi regressiokerroin).

Merkitään Y:llä osittain havaittua satunnaisotosta multinomijakaumasta $P(Y|\theta)$. Oletetaan, että Y:n jakauma voidaan täydellisesti määritellä parametrin θ avulla. Parametrin θ määrittelyä varten voidaan muodostaa Gibbsin otantaa (Gibbs sampler) muistuttava algoritmi, jota käytetään parametrien estimoimiseen.

Ideana otannassa on, että aluksi arvataan jotkut alkuarvot, joiden pohjalta lähdetään iteroimaan niin kauan, kunnes ehdolliset jakaumat suppenevat kohti jotakin jakaumaa.

Tämä algoritmi toimii MICE:n taustalla. MICE-algoritmin avulla saadaan θ:n posteriorijakauma ottamalla otoksia iteratiivisesti ehdollisista jakaumista

$$\begin{aligned} &P(Y_1|Y_{-1}, \theta_1) \\ &\quad\vdots \\ &P(Y_p|Y_{-p}, \theta_p). \end{aligned}$$

Parametrit $\theta_1, ..., \theta_p$ liittyvät vastaaviin ehdollisiin tiheyksiin. Nyt, kun havaituista reunajakaumista otetaan satunnaisotos, niin t:s iteraatio ketjutetuista yhtälöistä on Gibbsin otanta seuraavasti:

$$\begin{aligned} \theta_1^{*(t)} &\sim P(\theta_1|Y_1^{obs}, Y_2^{t-1}, ..., Y_p^{(t-1)}) \\ Y_1^{*(t)} &\sim P(Y_1|Y_1^{obs}, Y_2^{(t-1)}, ..., Y_p^{t-1}, \theta_1^{*(t)}) \\ &\quad\vdots \\ \theta_p^{*(t)} &\sim P(\theta_p|Y_p^{obs}, Y_2^{t-1}, ..., Y_p^{(t-1)}) \\ Y_p^{*(t)} &\sim P(Y_p|Y_p^{obs}, Y_2^{(t-1)}, ..., Y_p^{t-1}, \theta_p^{*(t)}) \end{aligned}$$

missä $Y_j^t = (Y_j^{obs}, Y_j^{*(t)})$ on j:nnes imputoitu muuttuja t:nnellä iteraatiolla. Tässä parametrin arvo muuttuu tarkemmaksi iteraatio iteraatiolta, kun kaikkia muita arvoja käytetään selittäjinä, ja selittäjien määrä kasvaa siten jokaisella iteraatiolla. Aiemmat imputoinnit $Y_j^{*(t-1)}$ kuuluvat $Y_j^{*(t)}$:hin ainoastaan suhteessa muihin muuttujiin, eivätkä suoraan. Tämän vuoksi parametrin arvot voivat supeta oikeita arvoja kohti hyvinkin nopeasti. Van Buurenin ja Groothuis-Oudshoornin (2011) mukaan 10-20 iteraatiota on usein riittävä määrä. [van Buuren and Groothuis-Oudshoorn, 2011]

Mallia on kritisoitu muun muassa siitä, ettei ehdollisessa imputoinnissa käytettävä algoritmi ole todellisuudessa Gibbsin otanta, minkä vuoksi ehdolliset jakaumat eivät välttämättä ole yhteensopivia todellisen yhteisjakauman kanssa. Jos kuitenkin ehdollisten jakaumien määrittämä yhteisjakauma on olemassa, tällöin kyseessä todella on Gibbsin otanta. [van Buuren, 2007] Ehdollisessa imputoinnissa muodostuva jakauma ei kuitenkaan välttämättä suppene oikeaan posteriorijakaumaan. Esimerkiksi Li, Yu ja Rubin (2012) ovat osoittaneet, että muuttujien imputointijärjestys vaikuttaa siihen, mitä jakaumaa kohti ehdolliset jakaumat suppenevat. Tämän vuoksi he ovat ehdottaneet ehdollisessa moni-imputoinnissa käytettävää algoritmiä kutsuttavan mahdollisesti epäsopivaksi Gibbsin otannaksi (possibly incompatible Gibbs Sampler, PIGS). [Kropko et al., 2014]

3.4.2 Yhdistettyjen mallien imputointi (JM)

Yhdistettyjen mallien imputointi olettaa mallin muuttujien noudattavan yhdessä multinormaalijakaumaa. Normaalijakaumaoletus ei kuitenkaan aina päde, erityisesti datan sisältäessä kaksiluokkaisia tai kategorisia muuttujia. Kuitenkin mm. [Schafer, 1997] ehdottaa, että multinormaalijakaumaoletus voi tulla oikein sovellettuna kysymykseen jopa kategoristen muuttujien kohdalla. Menetelmää onkin sovellettu paljon myös tilanteissa, joissa selkeästi nähdään, ettei multinormaalijakaumaoletus pidä paikkaansa.

Tästä lähin tässä tekstissä yhdistettyjen mallien imputoinnista käytetään nimitystä JM-moni-imputointi (joint modeling). JM-moni-imputointi voidaan pohjata EM-algoritmiin (Expectation Maximization = odotusten maksimointi). EM-algoritmi on samantapainen kuin Gibbsin otanta -algoritmi, mutta kun Gibbsin otanta on bayesiläinen, niin EM-algoritmi on frekventistinen.

Ennen EM-algoritmia aineisto jaetaan osiin bootstrap-otannan avulla. Bootstrap-otannassa aineistosta poimitaan otoksia takaisinpanolla. Jos oletetaan, että $S = \{i : 1, ..., n\}$ on otos toisistaan riippumattomia havaintoja ja $S^{(b)}$ on n:n havainnon otos S:stä, niin bootstrap-otokset saadaan asettamalla paino $m_i^{(b)}$ havainnolle i, missä

$$(m_1^{(b)}, ..., m_n^{(b)}) \sim MNOM[n; (n^{-1}), (n^{-1}), ..., (n^{-1})]$$

noudattaa multinomijakaumaa otoskoolla n ja n:llä solulla ja jokaisella yhtä suuri todennäköisyys $1/n$. $m_i^{(b)}$ on se lukumäärä, kuinka monta kertaa havainto i tulee valituksi b:nteen bootstrap-otokseen, $\sum_{i=1}^{n} m_i^{(b)} = n$.
Tämän jälkeen osille suoritettava EM-algoritmi muotoilee *ad hoc* -menetelmän, jonka avulla imputoida aineiston puuttuvuutta:

- Ensin puuttuvat arvot täydennetään estimoiduilla arvoilla.
- Tämän jälkeen estimoidaan parametrit.
- Tämän jälkeen estimoidaan puuttuvia arvoja uudelleen olettaen, että uudet parametriestimaatit ovat oikein.
- Lopuksi parametrit estimoidaan uudelleen.

Näitä vaiheita toistetaan iteratiivisesti yhä uudelleen suppenemiseen saakka.

Jokainen iteraatio koostuu E- ja M-vaiheista. E viittaa odotusarvoon (expectation) ja M maksimointiin (maximization). M-vaihe on yksinkertaisesti parametrin θ suurimman uskottavuuden estimointia olettaen, että puuttuvaa dataa ei ole lainkaan. Vaihe on siis laskennallisesti identtinen $l(\theta|Y)$:n tavallisen suurimman uskottavuuden estimoinnin kanssa.

E-vaiheessa muodostetaan ehdollinen odotusarvo puuttuvasta aineistosta ehdolla havaittu aineisto ja estimoidut parametrit ja puuttuvuutta imputoidaan näillä odotuksilla. EM-algoritmin perusideana on, että puuttuvuutta eivät olekaan Y_{mis}:t vaan Y_{mis}:n funktiot täydellisen datan log-uskottavuudessa $l(\theta|Y)$.

Oletetaan, että $\theta^{(t)}$ on parametri θ:n tämänhetkinen estimaatti. Tällöin EM-algoritmin vaihe E laskee täydellisen datan log-uskottavuuden seuraavasti:

$$Q(\theta|\theta^{(t)}) = \int l(\theta|y) f(Y_{mis}|Y_{obs}, \theta = \theta^{(t)}) dY_{mis}.$$

M-vaihe vastaavasti määrittää $\theta^{(t+1)}$:n maksimoimalla edellä lasketun täydellisen aineiston log-uskottavuuden seuraavasti:

$$Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta|\theta^{(t)})$$

kaikille θ. [Schafer, 1997]

3.4.3 Ehdollisen moni-imputoinnin ja JM-moni-imputoinnin vertailua

JM-moni-imputoinnilla ja ehdollisella moni-imputoinnilla on muutamia keskeisiä yhtäläisyyksiä: Molemmissa pyritään yhteisjakaumaan, josta imputointiarvot poimitaan. Lisäksi molemmille on keskeistä juuri moni-imputointi, koska se mallintaa imputoinnista syntyvää epävarmuutta. [Kropko et al., 2014]

JM-moni-imputoinnilla ja ehdollisella moni-imputoinnilla on myös yhteys. Merkitään $X = (X_1, ..., X_l)$ täydellisesti havaittuja kovariaattimuuttujia. Nyt jos $P(X, Y)$ on multinormaalisti jakautunut, niin kaikki ehdolliset tiheydet ovat lineaarisia regressioita, joilla on vakio, normaalisti jakautunut virhevarianssi. Tällöin myös ehdollinen jakauma $P(Y_j|X, Y_{-j})$ noudattaa lineaarista regressiomallia. Nyt, jos tällaisessa tilanteessa lähdetään suorittamaan ehdollista moni-imputointia, on se identtinen JM-moni-imputoinnin kanssa. Tällöin myöskin ehdollisessa moni-imputoinnissa käytettävä algoritmi on Gibbsin otanta, koska suppeneminen kohti oikeaa jakaumaa tapahtuu.

Monissa tutkimuksissa, esim. [Kropko et al., 2014], [van Buuren, 2007] ja [Lee and Carlin, 2010] on todettu ehdollisen moni-imputoinnin olevan joustavampi menetelmä kuin JM-moni-imputointi, koska se mahdollistaa suuremman erilaisten jakaumien kirjon kuin JM. Ehdollisessa moni-imputoinnissa imputointimallit määritellään jokaiselle muuttujalle erikseen, joka mahdollistaa erilaisten muuttujatyyppien huomioinnin: Esimerkiksi kaksiluokkaista muuttujaa voidaan imputoida logistisella regressiolla ja useampiluokkaista muuttujaa voidaan imputoida usean luokan logistisella regressiolla.

JM-moni-imputointia ja ehdollista moni-imputointia vertailevissa artikkeleissa on saatu toisistaan poikkeavia tuloksia. Esimerkiksi [van Buuren, 2007] osoittavat artikkelissaan, että jos aineisto ei ole multinormaalisti jakautunut, ei multinormaalioletus heijastele todellista jakaumaa ja ehdollinen moni-imputointi olisi siten järkevämpi vaihtoehto. Samaan tulokseen päätyy myös [Kropko et al., 2014], vaikka artikkelissa todetaankin sen tarjoavan vain yhden näkökulman aiheeseen. Sen sijaan [Lee and Carlin, 2010] esittää normaalioletuksen toimivan hyvin. He jopa toteavat normaalioletuksen toimivan ehdollista moni-imputointia paremmin kaksiluokkaisille muuttujille. Näiden artikkelien perusteella vaikuttaisi siis siltä, että kumpaakaan näistä menetelmistä ei voida valita absoluuttisesti paremmaksi kuin toista, vaan niiden hyvyys riippuu mm. kulloinkin käytetystä aineistosta ja imputoidun aineiston käyttötarkoituksesta.

3.5 Muita mahdollisia menetelmiä

Puuttuvia arvoja sisältävää dataa voidaan hyödyntää myös suorittamatta imputointia. Yksinkertaisin ja yleisin tapa on jättää analyysin ulkopuolelle ne havainnot, joilta ei ole saatu arvoja kiinnostuksen kohteena oleviin muuttujiin. Tämä menetelmä sisältää kuitenkin riskejä, erityisesti sen vuoksi, että useinkaan puuttuvuus ei ole täysin satunnaista. Havainnot eivät ole tällöin satunnaisotos alkuperäisestä otoksesta, jolloin tällaisesta datasta lasketut estimaatit ovat harhaisia. Puuttuvien arvojen jättämistä pois data-analyysissä kutsutaan nimellä complete case analysis (CC). Tästä käytetään myös nimeä listapoistaminen (listwise deletion). Mikäli puuttuvuus on MCAR, eli täysin satunnaista, voidaan puuttuvat havainnot yleensä jättää aineistosta pois ilman, että siitä koituu huomattavaa haittaa. Tällainen tilanne on kuitenkin erittäin harvinainen survey-tutkimuksissa.

Parittaisesta poistamisesta (pairwise deletion) on kyse silloin, kun yksittäisen muuttujan kaikkia saapuvilla olevia arvoja käytetään. Tämä voikin olla tietyssä määrin toimiva menetelmä yksittäisen muuttujan tarkasteluun, mutta ongelmia voi syntyä, jos aletaan tarkastelemaan esimerkiksi muuttujien välisiä suhteita. Yksittäisten muuttujien estimaa-

tit oltaisiin laskettu kaikista kyseisten muuttujien havaituista arvoista, mutta esimerkiksi korrelaation laskemiseen otettaisiin mukaan vain ne rivit, joissa kumpikin kahdesta muuttujasta on havaittu. [Pigott, 2001]

Painotusmenetelmiä voidaan käyttää puuttuvien arvojen paikkaamiseen analyysissä siten, että joillekin havaituille muuttujan arvoille annetaan enemmän painoarvoa kuin toisille. Jokaiselle havainnolle annetaan oma painokerroin, jolla muuttujan arvoa painotetaan. [Pentala, 2014]

Jos tehdään pitkittäistutkimusta, voidaan puuttuva arvo imputoida myös aiemmalla tutkimuskerralla saadulla arvolla. Tämä menetelmä toimii erityisesti silloin, jos tutkimuksen kuluessa kyseisen arvon ei odoteta muuttuvan. Tällainen muuttuja voisi olla esimerkiksi sisarusten lukumäärä. Sen sijaan, jos imputoidaan analyysin kohteena olevaa muuttujaa, voi tällainen imputointi vääristää tuloksia, aliarvioimalla ennustavien tekijöiden vaikutusta muuttujan arvon muutokseen. [Molnar et al., 2008]

Luku 4

Esimerkki: GeneRISK-tutkimuksen kyselyaineisto

Sydän- ja verisuonitaudit ovat yleisimpiä suomalaisia kansantauteja, sekä suomalaisten yleisin kuolinsyy. Perinteiset riskitekijät, kuten elintavat, selittävät noin puolet sairastumisriskin vaihtelusta, loppu liittyy perinnöllisiin riskeihin. Nämä seikat yhdessä ovat keskeisiä syitä sille, miksi tutkimusta sydän- ja verisuonitaudeista tehdään jatkuvasti ja monista eri lähestymiskulmista. Nykyisellään moni korkean sydäntautiriskin omaava henkilö jää huomaamatta, koska kaikkia riskitekijöitä ei tunneta.

Tutkielmassa käytetään GeneRISK-tutkimuksessa kerättyä aineistoa. GeneRISK-tutkimus pyrkii kartoittamaan sitä, voitaisiinko sydän- ja verisuonitauteja ehkäistä, jos saatavilla on sellaista tietoa perimästä, joka vaikuttaa kyseisten sairauksien syntyyn. Tutkimus on osa SalWen koordinoimaa Yksilöllistetty diagnostiikka ja hoito (GET IT DONE) –ohjelmaa. [GeneRISK-verkkosivut, c] Sairastumisriskin arvio perustuu tilastolliseen mallinnukseen, jota on käytetty 19000 suomalaista käsittävässä, vuosina 1992, 1997, 2002 ja 2007 kerätyssä FINRISKI-aineistossa. [GeneRISK-verkkosivut, b]

Tavoitteena tutkimuksessa on yksilön perimästä saatavan tiedon hyödyntäminen siten, että siitä olisi apua niin terveydenhuollossa ja sairaanhoidossa kuin yksilölähtöisessä sai-

rauden ennaltaehkäisyssä. Geeniprofiilitieto voisi tuoda merkittävän lisän sydän- ja verisuonitautien riskinarvioon. Ennaltaehkäisy säästäisi myös terveydenhuollon kustannuksia, joista suurin osa tulee sairauksien hoidosta.

Myös yksilöiden terveyskäyttäytyminen riskin saamisen jälkeen on GeneRISK-tutkimuksen kiinnostuksen kohteena. Tätä tutkitaan kutsumalla tutkittavia seurantatutkimuksiin. Kyseessä on ensimmäinen vastaavassa mittakaavassa toteutettu tutkimus, jossa geeniprofiilitieto palautetaan tutkittaville. Tutkimuksen avulla pyritään viemään sairauksien hoitoa ennaltaehkäisevään suuntaan. [GeneRISK-verkkosivut, g] Tutkimuksessa kerätyt tiedot talletetaan Terveyden ja hyvinvoinnin laitoksen (THL) biopankkiin, sekä Carean tapauksessa sydän- ja verisuonitapahtuman riskiarvio lisäksi Carean järjestelmän potilastietoihin. [GeneRISK-verkkosivut, e]

4.1 Terveystietokyselyistä yleisesti

Terveystietokyselyaineistoa, jollainen myös tutkielmassa käytetty aineisto on, tutkittaessa on syytä ottaa huomioon joitakin seikkoja, jotka saattavat vaikuttaa aineiston luotettavuuteen. Kyselytutkimukseen saattaa vaikuttaa häiriötekijöitä, kuten vastaustilanne tai yhteiskunnallinen mielipideilmasto. Myös eri väestöryhmät saattavat vastata kysymyksiin eri tavoin: Esimerkiksi elintapoja raportoitaessa saatetaan tietoja vähätellä tai liioitella sen mukaan, mikä katsotaan sosiaalisesti suotavaksi. Lisäksi esimerkiksi vastaajan äidinkieli saattaa vaikuttaa vastauksien todenmukaisuuteen, mikäli kysymystä ei ymmärretä oikein. Tutkimuslomakkeessa on pyrittävä rajaamaan kysymykset tarkasti ja ilmaisemaan ne selkeästi, jotta väärinymmärryksiltä vastattaessa vältyttäisiin. [Tiirikainen, 2006]

Terveystietokyselyyn kutsutaan vastaajia usein henkilökohtaisesti. Voitaisiin ajatella, ettei tutkimuksesta tällöin kieltäydytä yhtä herkästi kuin esimerkiksi satunnaisesta puhelinhaastattelusta. Kuitenkin survey-tutkimusta tehtäessä on otettava huomioon myös se, että tutkimukseen osallistuminen on vapaaehtoista, jolloin kutsuttujen joukossa on aina

niitä, jotka kieltäytyvät osallistumasta. Ei voida myöskään sulkea pois sitä mahdollisuutta, että tutkimuksesta kieltäytyminen olisi jollain tavoin systemaattista jonkin tietyn väestöryhmän keskuudessa. Esimerkiksi GeneRISK-tutkimuksessa ei voida varmuudella tietää kaikkia niitä syitä, jotka saavat kutsuttuja kieltäytymään osallistumasta.

4.2 Aineiston keruu

Aineisto käsittää 1400 havaintoa Kymenlaakson sairaanhoito- ja sosiaalipalvelujen kuntayhtymä Carean aineistosta (N=1400, tilanne 18.1.2016). Tästä poistetaan vielä rivit, joilta puuttuu havaintoarvoja myöhemmin esiteltäviin taustamuuttujiin. Tämän jälkeen lopullisen tutkielmassa käytettävän aineiston N=1278. GeneRISK-tutkimuksen tutkittavien rekrytointi alkoi maaliskuussa 2015 [GeneRISK-verkkosivut, a] ja pilotointivaihe päättyi loppuvuodesta 2015. Tutkittavien rekrytointi jatkuu syksyyn 2017 saakka, mutta tutkimuksen kokonaiskesto on 20 vuotta, mikä mahdollistaa tilanteen seuraamisen vielä pitkään tiedon keräämisen jälkeen. Pitkäaikaisella tutkimuksella voidaan ennustaa riskitekijöitä, sekä ymmärtää tautien taustoja. [GeneRISK-verkkosivut, h]

Kysely ei ole perinteinen survey-tutkimus siinä mielessä, että kyselylomake on vain yhdessä roolissa koko tutkimuksessa. Carea kutsuu tutkimukseen kunnittain kaikki alueella asuvat 45-64-vuotiaat henkilöt, joilla ei vielä ole ollut sepelvaltimotapahtumaa, kunnes 5000 henkilöä on saatu rekrytoitua. Yksikkövastauskatoa ei siis pääse syntymään, toisin kuten monissa muissa survey-tutkimuksissa. [GeneRISK-verkkosivut, d] Etelä-Kymenlaakson alueella, josta tutkittavat rekrytoidaan, asuu 26 400 45-65-vuotiasta. (Tilastokeskus, 31.12.2014) Helmikuussa 2016 tutkimukseen osallistumisprosentti kutsuttujen joukosta on arvioitu olevan n. 26 %.

Perustietokyselylomake, josta tässä tutkielmassa käytettävä aineisto on peräisin, on mahdollista täyttää joko sähköisessä muodossa tutkimusportaalissa (my.generisk.fi) tai paperisena, jolloin se täytyy erikseen pyytää. Vastaajilla on kuitenkin oltava internet-yhteys käytössään voidakseen osallistua tutkimukseen, joten suurin osa tässä tutkielmassa käy-

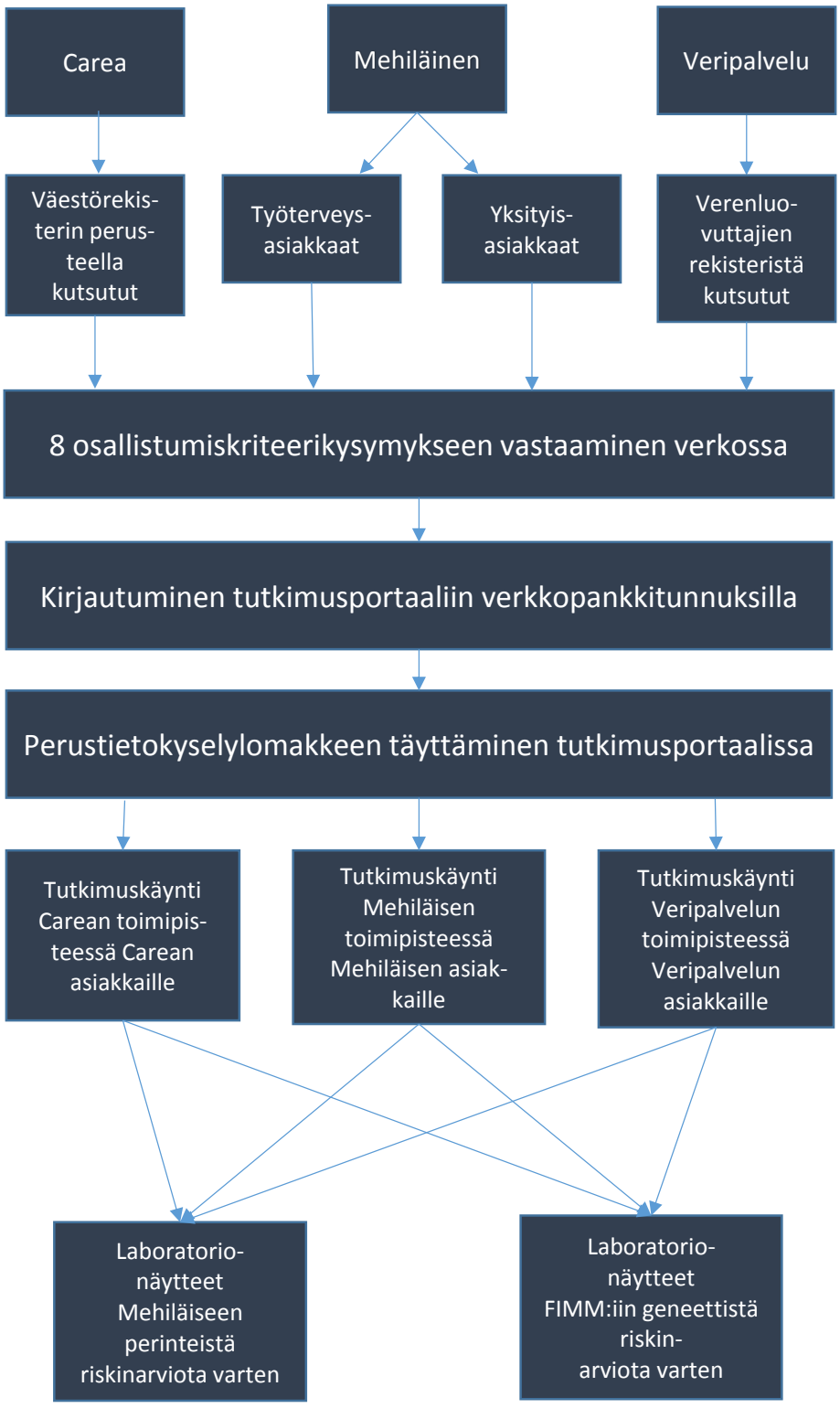
tettävän aineiston vastauksista on saatu sähköisessä muodossa. Aineistossa on kaksi henkilöä, jotka ovat täyttäneet lomakkeen paperisessa muodossa, josta tutkimushoitaja on sitten kirjannut vastaukset sähköiseen tutkimusportaaliin. Nämä henkilöt eivät ole jättäneet vastaamatta yhteenkään pakolliseen kysymykseen, joten vastaukset on voitu kirjata sähköiseen portaaliin ongelmitta. Mikäli näin ei olisi ollut, tutkimushoitajat olisivat olleet yhteydessä tutkittaviin henkilökohtaiseksi kyselläkseen lisätietoja.

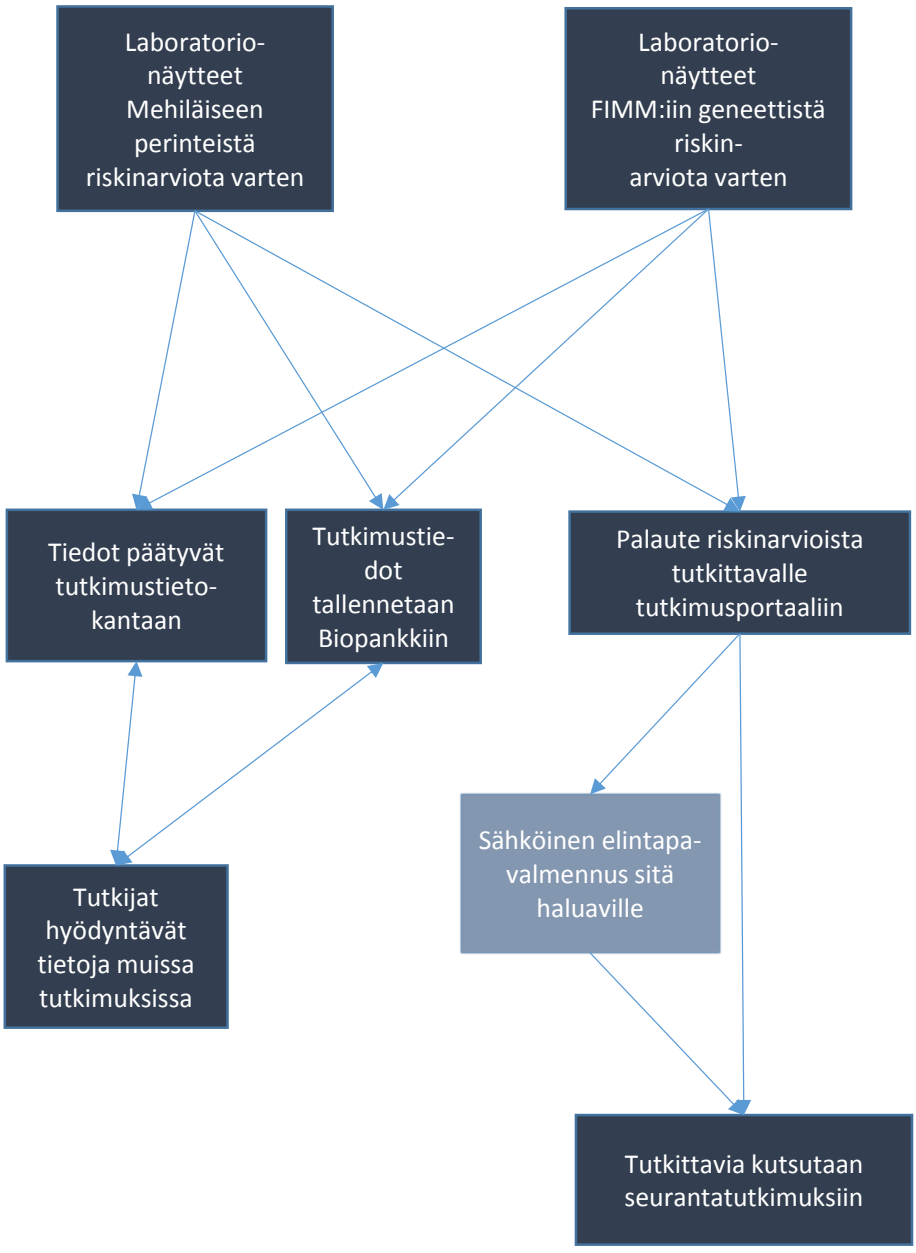
GeneRISK-tutkimuksen kiinnostusperusjoukkona on suomalainen 45-65 vuotias väestö. Näistä tutkimukseen on valittu tavoiteperusjoukoksi Veripalvelun verenluovuttajia, Mehiläisen työterveyshuollon asiakkaita työterveyshuollon vuositarkastusten yhteydessä, sekä Kymenlaakson sairaanhoito- ja sosiaalipalvelujen kuntayhtymä Carean alueella asuvia henkilöitä. [GeneRISK-verkkosivut, d] Tässä tutkielmassa käytetään vain Carean aineistoa.

4.3 Tutkimuksen vaiheet

GeneRISK-tutkimukseen rekrytoidaan osallistujia kolmesta eri keskuksesta: Mehiläisestä, Veripalvelusta ja Careasta. Veripalvelu kutsuu mukaan verenluovuttajia rekisteristään. Mehiläinen kutsuu työterveyden vuositarkastuksen yhteydessä asiakkaitaan mukaan, sekä avoimella kutsulla kaikkia yksityisasiakkaitaan. Carea kutsuu Kymenlaakson sairaanhoito- ja sosiaalipalvelujen kuntayhtymän alueelta tutkittavia väestörekisterin perusteella siten, että ensimmäisen vuoden aikana tutkimukseen on kutsuttu noin 42 % sopivan ikäisistä henkilöistä. Kutsuttaville lähetetään henkilökohtainen kutsukirje (Liite B), jossa kerrotaan tutkimuksesta, mm. sen tarkoituksesta, kulusta ja osallistumiskriteereistä.

Kutsun saatuaan kutsuttava rekisteröityy tutkimusportaaliin omilla verkkopankkitunnuksilla. Ennen rekisteröitymistä henkilön sopivuus tutkimukseen varmistetaan kyselyllä, jossa kysytään kahdeksan osallistumiskriteerien täyttymistä kartoittavaa kysymystä. (ks. luku 4.5.1) Tutkimuksen vaiheet on esitetty seuraavassa kaaviossa.


Carea
Mehiläinen
Veripalvelu
Väestörekis-terin perus-teella kutsutut
Työterveys-asiakkaat
Yksityis-asiakkaat
Verenluo-vuttajien rekisteristä kutsutut
8 osallistumiskriteerikysymykseen vastaaminen verkossa
Kirjautuminen tutkimusportaaliin verkkopankkitunnuksilla
Perustietokyselylomakkeen täyttäminen tutkimusportaalissa
Tutkimuskäynti Carean toimipis-teessä Carean asiakkaille
Tutkimuskäynti Mehiläisen toimipisteessä Mehiläisen asiak-kaille
Tutkimuskäynti Veripalvelun toimipisteessä Veripalvelun asiakkaille
Laboratorio-näytteet Mehiläiseen perinteistä riskinarviota varten
Laboratorio-näytteet FIMM:iin geneettistä riskin-arviota varten


Laboratorio-
näytteet
Mehiläiseen
perinteistä
riskinarviota varten
Laboratorio-
näytteet
FIMM:iin geneettistä
riskin-
arviota varten
Tiedot päätyvät
tutkimustieto-
kantaan
Tutkimustie-
dot
tallennetaan
Biopankkiin
Palaute riskinarvioista
tutkittavalle
tutkimusportaaliin
Sähköinen elintapa-
valmennus sitä
haluaville
Tutkijat
hyödyntävät
tietoja muissa
tutkimuksissa
Tutkittavia kutsutaan
seurantatutkimuksiin

4.4 Aineiston muuttujat

Useimmat aineiston muuttujat ovat diskreettejä, eli epäjatkuvia. Aineisto sisältää luokitteluasteikollisia, järjestysasteikollisia, suhdeasteikollisia ja jatkuvia muuttujia.

Luokitteluasteikolliset muuttujat ovat sellaisia, joita ei voi asettaa keskenään mihinkään suuruus- tai paremmuusjärjestykseen. Tällainen muuttuja aineistossa on esimerkiksi ”Mistä tupakointikerrasta teidän on vaikeinta luopua?” Kysymyksen vastausvaihtoehdot ovat ”Aamun ensimmäisestä” ja ”Jostain muusta”. Tässä tapauksessa kyse on itse asiassa vielä kaksiluokkaisesta, eli dikotomisesta muuttujasta.

Järjestysasteikolliset muuttujat ovat sellaisia, joiden arvot voidaan asettaa johonkin järjestykseen, mutta arvojen väliset etäisyydet eivät ole välttämättä yhtä pitkät tai etäisyyksiä ei ole mielekästä tarkastella. Tällainen muuttuja aineistossa on esimerkiksi ”Kuinka tärkeitä seuraavat terveyteen vaikuttavat seikat ovat teille: Painonhallinta”. Kysymyksen vastausvaihtoehdot ovat ”Ei lainkaan tärkeä”, ”Vähän tärkeä”, ”Kohtalaisen tärkeä”, ”Melko tärkeä” ja ”Erittäin tärkeä”.

Lomakkeen kysymys ”Kuinka suuret olivat taloutenne (ruokakuntanne) kokonaistulot viime vuonna (veroja vähentämättä)?” Kysymyksessä on 12 eri vaihtoehtoa, joista viimeinen vaihtoehto on ”En halua vastata”, joka ei sovi suhdeasteikolle. Muut vaihtoehdot ovat rahasummien välejä, esimerkiksi vastausvaihtoehto 3 ”25001-35000 €”. Asteikolla on selkeä absoluuttinen nollakohta, eli 0 euroa, ja vaihtoehtojen väliset suhteet ovat yksikäsitteisesti määritettävissä.

Jatkuva muuttuja voi saada mitä tahansa arvoja tietyltä väliltä. Lomakkeen kysymykset, jotka alkavat ”Kuinka monta minuuttia istutte keskimäärin arkipäivänä...” ovat jatkuvia muuttujia, koska vastaukset tallennetaan minuutteina.

4.4.1 Väestöryhmään liittyvät muuttujat

Vastaamisinnokkuuteen ja vastauksiin saattavat vaikuttaa erityisesti väestöryhmiin liittyvät muuttujat. Väestöryhmiin liittyviä muuttujia voidaan käyttää myös esimerkiksi löytämään keskenään samankaltaisia henkilöitä imputointitarkoituksiin. Taustamuuttujien avulla tehtävässä imputoinnissa olisi ideaalista, ettei itse taustamuuttujissa esiintyisi puuttuvuutta, ellei johonkin taustamuuttujakysymykseen vastaamattomuutta itsessään pidetä kannanottona ja täten merkityksellisenä. Tässä aineistossa väestöryhmään liittyviä taustamuuttujia ovat:

- Sukupuoli (gender)
- Ikä (lasketaan syntymävuodesta = ageofbirth)
- Siviilisääty (15.5)
- Asuinalue (15.1)
- Syntymäkotikunta (15.2)
- Sosioekonominen asema (15.11)
- Koulutus (15.6)
- Kotitalouden tulot (15.12)

Tätä tarkastellessa huomattiin, että myös taustamuuttujat sisältävät melkoisesti puuttuvuutta. Ainoastaan ikä, sukupuoli ja syntymäkotikunta eivät sisällä lainkaan puuttuvuutta: Ne ovat pakollisia tietoja kaikille osallistujille, joten siksi puuttuvuutta ei ole.

Runsaan puuttuvuuden vuoksi päädyttiin aineistosta poistamaan kaikki sellaiset rivit, joilla kaikki tai osa kaikkien näiden taustamuuttujien, paitsi ”Kotitalouden tulot” -muuttujan arvoista on puuttuvia. Näin kyseisiä muuttujia voidaan helpommin käyttää taustamuuttujina esiteltäessä puuttuvuutta ja aineiston tilastollisen täydentämisen menetelmiä. Kotitalouden tuloja sisältävästä muuttujasta poistettiin muu puuttuvuus, mutta ”En halua vastata” vaihtoehto muutettiin puuttuvuudeksi tämän muuttujan kohdalla. Tämä aineiston rivien poisto on tehtiin havainnollisuuden vuoksi, eikä se ole suositeltavaa kun tilastollista täydentämistä lähdetään tekemään tositilanteessa.

4.4.2 Elintapoihin liittyvät muuttujat

Elintapoihin liittyvät muuttujat ovat sellaisia muuttujia, jotka vaikuttavat taustamuuttujien ohella tautiriskeihin. Toisin kuin usein väestöryhmiin liittyviin muuttujiin, näiden arvoihin tutkittava voi itse vaikuttaa. Kyselylomakkeessa on mm. seuraaviin elämäntapavalintoihin liittyviä kysymyssarjoja:

- Tupakointi
- Alkoholinkäyttö
- Liikunta
- Ruokailu

Näistä tähän tutkielmaan kiinnostuksen kohteena oleviksi muuttujiksi valikoitiin seuraavat liikuntaan ja ruokailuun liittyvät muuttujat:

Liikunta

1. Kuinka paljon keskimäärin liikut ja rasitat itseäsi ruumiillisesti vapaa-aikana? (5.2) (arvot 1,2,3,4,5, joista 1 kuvaa vähäisintä mahdollista liikuntaa ja 5 aktiivista kilpaurheilua)
2. Kuinka monta tuntia istutte keskimäärin arkipäivänä? työpäivän aikana toimistossa tai vastaavassa (5.4A) (aika minuutteina)
3. Kuinka monta tuntia istutte keskimäärin arkipäivänä? kotona televisiota tai videoita katsellen (5.4B) (aika minuutteina)
4. Kuinka monta tuntia istutte keskimäärin arkipäivänä? kotona tietokoneen ääressä (5.4C) (aika minuutteina)
5. Kuinka monta tuntia istutte keskimäärin arkipäivänä? kulkuneuvossa (5.4D) (aika minuutteina)
6. Kuinka monta tuntia istutte keskimäärin arkipäivänä? muualla (5.4E) (aika minuutteina)

Istumista koskevat muuttujat yhdistettiin summamuuttujaksi (istu), johon kaikki istumista koskevat muuttujat laskettiin yhteen. Jos istumisen minuuttimäärä ylitti yhden vuorokauden, havainto muutettiin puuttuvaksi.

Ruokailu

1. Kuinka monena arkipäivänä viikossa syötte seuraavat pääateriat? Aamupala (11.1A) (0 = ”En syö”, 1 = ”1-2 päivänä”, 2 = ”3-4 päivänä”, 3 = ”Joka päivä”)
2. Kuinka monena arkipäivänä viikossa syötte seuraavat pääateriat? Lounas (11.1B) (0 = ”En syö”, 1 = ”1-2 päivänä”, 2 = ”3-4 päivänä”, 3 = ”Joka päivä”)
3. Kuinka monena arkipäivänä viikossa syötte seuraavat pääateriat? Päivällinen/iltaruoka (11.1C) (0 = ”En syö”, 1 = ”1-2 päivänä”, 2 = ”3-4 päivänä”, 3 = ”Joka päivä”)
4. Kuinka usein tavallisesti käytätte seuraavia elintarvikkeita: ruis-, kokojyvä- tai näkkileipä (11.3A) (arvot 1,2,...,8, joista 1 = ”Harvemmin kuin kerran kuukaudessa tai ei lainkaan” ja 8 = ”4 kertaa päivässä tai useammin”)
5. Kuinka usein tavallisesti käytätte seuraavia elintarvikkeita: tuoresalaattia tai muita kasviksia (11.3G) (arvot 1,2,...,8, joista 1 = ”Harvemmin kuin kerran kuukaudessa tai ei lainkaan” ja 8 = ”4 kertaa päivässä tai useammin”)
6. Kuinka usein tavallisesti käytätte seuraavia elintarvikkeita: keitettyjä lisäkekasviksia tai palkokasviksia (11.3H) (arvot 1,2,...,8, joista 1 = ”Harvemmin kuin kerran kuukaudessa tai ei lainkaan” ja 8 = ”4 kertaa päivässä tai useammin”)
7. Kuinka usein tavallisesti käytätte seuraavia elintarvikkeita: hedelmiä tai marjoja (11.3I) (arvot 1,2,...,8, joista 1 = ”Harvemmin kuin kerran kuukaudessa tai ei lainkaan” ja 8 = ”4 kertaa päivässä tai useammin”)
8. Kuinka usein tavallisesti käytätte seuraavia elintarvikkeita: kalaa tai kalaruokia (11.3N) (arvot 1,2,...,8, joista 1 = ”Harvemmin kuin kerran kuukaudessa tai ei lainkaan” ja 8 = ”4 kertaa päivässä tai useammin”)
9. Kuinka usein tavallisesti käytätte seuraavia elintarvikkeita: pikaruokia (esim. hampurilainen tai pizza) (11.3O) (arvot 1,2,...,8, joista 1 = ”Harvemmin kuin kerran kuukaudessa tai ei lainkaan” ja 8 = ”4 kertaa päivässä tai useammin”)
10. Kuinka usein tavallisesti käytätte seuraavia elintarvikkeita: voita, voi-kasviöljyseosta (esim. Oivariini, Ingmariini) (11.3R) (arvot 1,2,...,8, joista 1 = ”Harvemmin kuin kerran kuukaudessa tai ei lainkaan” ja 8 = ”4 kertaa päivässä tai useammin”)
11. Kuinka usein tavallisesti käytätte seuraavia elintarvikkeita: makeita leivonnaisia, jälkiruokia, makeisia tai suklaata (11.3T) (arvot 1,2,...,8, joista 1 = ”Harvemmin

kuin kerran kuukaudessa tai ei lainkaan" ja 8 = "4 kertaa päivässä tai useammin")

12. Kuinka usein tavallisesti käytätte seuraavia elintarvikkeita: sokeroituja virvoitusjuomia tai mehuja (11.3U) (arvot 1,2,...,8, joista 1 = "Harvemmin kuin kerran kuukaudessa tai ei lainkaan" ja 8 = "4 kertaa päivässä tai useammin")

4.5 Yksikkövastauskadon tarkastelu

Yksikkövastauskadolla tarkoitetaan sitä, kun joltain tutkimukseen kutsutulta ei ole saatu lainkaan vastauksia tai ne ovat käyttökelvottomia. Tätä voidaan tarkastella esimerkiksi otostiedostojen ja rekisterien avulla, jos tällaisia on saatavilla. [Pentala, 2014] Tutkielmassa käytettävän tutkimusaineiston osalta on tiedossa osallistumisprosentti, eli se, kuinka moni tutkimukseen kutsutuista on päättänyt osallistua tutkimukseen. Käytettävissä ei kuitenkaan ollut rekisteritietoa niistä kutsutuista, jotka kieltäytyivät osallistumasta. Osallistumisinnokkuutta joudutaan siis tarkastelemaan suppeampien tietojen avulla.

4.5.1 Osallistumiskriteerit

Tutkimukseen osallistumiseen vaikuttavat monet seikat. Tutkittaville on asetettu muutamia osallistumiskriteerejä, jotka ovat edellytyksenä tutkimukseen osallistumiselle. Nämä kahdeksan kriteeriä on listattu seuraavaksi.

1. Tutkittavan on oltava 45-65 –vuotias.
2. Tutkittava ei saa olla sairastanut sydänveritulppaa (sydäninfarktia).
3. Tutkittava ei saa olla sairastanut aivohalvausta, aivoverenvuotoa tai aivoverisuonitukosta.
4. Tutkittava ei saa sairastaa sepelvaltimotautia.
5. Tutkittavalla ei saa olla taustalla sepelvaltimon (sydän) ohitusleikkausta.
6. Tutkittavalla ei saa olla taustalla sepelvaltimon (sydän) pallolaajennusta.

7. Tutkittavalla ei saa olla määrättyä edunvalvojaa.

8. Tutkittava ei saa olla raskaana.

Tällaiset tutkimuksessa asetetut vaatimukset sulkevat osallistujia tutkimuksen ulkopuolelle, vaikka halua osallistua olisikin. [GeneRISK-verkkosivut, d] Kutsuja lähetettäessä ei oteta huomioon näitä kahdeksaa osallistumiskriteeriä, joten ne ovat osaltaan syynä vastaamattomuuteen.

Survey-tutkimuksen laatua mitataan muun muassa vastausaktiivisuudella, ja se vaikuttaakin tulosten oikeellisuuteen ja yleistettävyyteen. [Tiirikainen, 2006] Esimerkiksi vastausten jakauma saattaa tällöin olla virheellisesti vinoutunut. Tutkimus voi vetää puoleensa tietynlaisia henkilöitä, jolloin vastaukset ovat keskenään samanlaisempia kuin muuten.

Poissulkukriteereiden lisäksi syynä vastaamattomuuteen voisi ajatella olevan esimerkiksi mahdollinen huono terveydentila: Jos henkilö joutuu viettämään paljon aikaa sairaalassa, tai hänen on vaikea poistua kotoaan, voi tutkimukseen osallistuminen olla tällöin ylivoimaista. Tutkimukseen vastaamatta jättäminen saattaisi johtua myös omista elintavoista: Kyselylomakkeessa kartoitetaan mm. nukkumista, liikuntatottumuksia, tupakan- ja alkoholinkäyttöä. Jos henkilö kokee, että hänen elintapansa eivät ole tarpeeksi hyvät, saattaa häpeä niistä johtaa siihen, ettei kyselyyn vastata lainkaan.

Tutkittavalla on oltava käytössään internet-yhteys, sekä sähköpostiosoite ja verkkopankkitunnukset, jotta tunnistautuminen ja tiedonkulku onnistuu. Tutkimuksessa on käytössä verkkoportaali, jossa perustietokyselylomake täytetään ja johon tutkimustulokset ilmestyvät niiden valmistuessa. Jos internet-yhteyttä ei ole, osallistuminen ei ole mahdollista ja kutsuttu jättäytyy pois tutkimuksesta.

4.5.2 Infopuhelimen kirjanpito

Tutkimusta toteutettaessa on kerätty myös niin sanottua infopuhelinaineistoa. Infopuhelimeen soitetaan, jos halutaan esimerkiksi kysyä lisätietoa tutkimuksesta tai varata tai peruuttaa aikaa. Tässä on pidetty kirjaa Carean Infopuhelimeen soittaneista henkilöistä. Puhelimeen soitetaan pääasiassa terveystarkastusajan varaamiseksi, mutta myös muista syistä, kuten kysymysten herätessä tai osallistumista peruttaessa, tai siksi, ettei alun perinkään haluta tai voida osallistua.

Tutkittiin aikavälillä 21.5.2015-22.2.2016 raportoituja soittoja. Tänä aikana puheluita, joissa soittaja ilmaisee, ettei voi tai halua osallistua tutkimukseen, on vastaanotettu 177 kappaletta. Tarkemmat tiedot ovat nähtävissä taulukossa.

Taulukko 4.1: Infopuhelimen kirjanpito.

Poisjääntisyy	**Lukumäärä**
Internet-yhteyden, sähköpostiosoitteen tai verkkopankkitunnuksien puute	111
Poissulkukriteerinä oleva sairaus	34
Vapaita tutkimusaikoja vasta pitkän ajan kuluttua	8
Henkilö raskaana	1
Henkilöllä edunvalvoja	1
Muu syy	22
Yhteensä	**177**

Muihin syihin lukeutui myös tapaus, jossa kutsuttu oli muuttanut toiselle paikkakunnalle, eikä siten enää kuulunut tutkimuksen tavoiteperusjoukkoon.

Soittoja tarkasteltaessa on syytä ottaa huomioon, että kaikki kutsun saaneista eivät soita infopuhelimeen kertoakseen, etteivät voi tai halua osallistuvansa tutkimukseen, vaan

suurin osa tällaisista henkilöistä vain yksinkertaisesti jättää osallistumatta ilman mitään ilmoitusta. Kutsukirjeessä ei kuitenkaan suoraan mainita poissulkukriteerejä, eikä sitä, että internetyhteys, pankkitunnukset ja sähköpostiosoite on oltava, jotta tutkimukseen voisi osallistua. Tämä saattaa selittää näitä koskevien soittojen suurta osuutta.

Tutkittava saa milloin tahansa tutkimuksen aikana peruuttaa osallistumisensa tutkimukseen, mutta tässä tutkielmassa käsitellään tutkimusprosessin vaiheita ainoastaan kyselylomakkeen täyttöön asti.

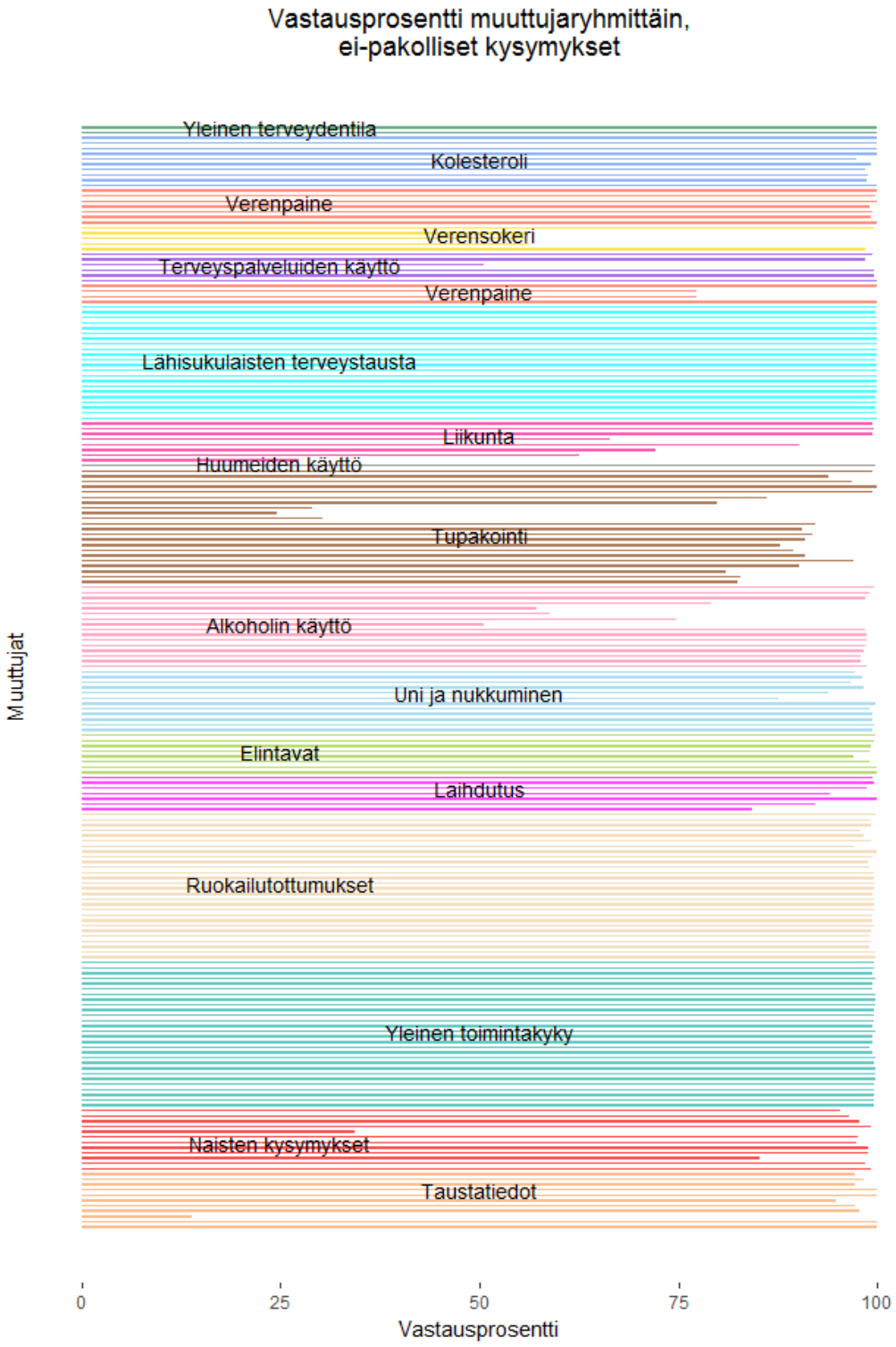
Luku 5

Analyysit ja tulokset

5.1 Erävastauskato tutkimusaineistossa

Tutkielmassa käytettävässä aineistossa esiintyy useissa muuttujissa erävastauskatoa (ks. 2.2). Puuttuvuutta rajoittaa kuitenkin esimerkiksi se, että verkkolomake estää tiettyjen kysymysten yli hyppäämisen, jolloin näihin on pakko vastata ennen seuraaviin kysymyksiin siirtymistä. Tällä pyritään varmistamaan, että tutkittavilta saadaan kaikkein olennaisimmat tiedot riskinarvion laskemiseksi. Muut kysymykset, jotka eivät ole pakollisia, antavat tarkentavia lisätietoja. Monipuolisen aineiston jatkokäytön kannalta olisi ihanteellista, että koko lomake olisi rakennettu siten, ettei kysymyksiä voi jättää välistä.

Tarkasteltiin yksitellen aineiston muuttujia puuttuvuuden näkökulmasta. Kuvassa 5.1 nähdään, millainen vastausprosentti on lomakkeen ei-pakollisissa kysymyksissä. Puuttuvuusprosentti on laskettu niistä henkilöistä, joille kysymys on osoitettu.



Kuva 5.1: Vastausprosentti muuttujaryhmittäin. Mukana vain ei-pakolliset kysymykset ja vastausprosentti laskettu niistä henkilöistä, joille kysymys on osoitettu.

Huom. Kysymys 5.4F ei löydy aineistosta, vaikka se löytyy kysymyslomakkeelta.

Taulukosta B huomataan, että datassa on kaksi kaikille tarkoitettua kysymystä, joissa puuttuvuus on selkeästi ylitse muiden. Kysymyksessä

- 2.3 ”Kuinka monta kokonaista päivää olitte viimeksi kuluneen vuoden (12 kk) aikana sairauden takia poissa töistä tai hoitamatta tavallisia tehtäviänne? (Jos ette yhtään, vastatkaa 0.)”

puuttuvia arvoja oli 1101 ja kysymyksessä

- 15.10 ”Kuinka moni taloutenne jäsenistä on 0-13v?”

631 puuttuvaa arvoa. Näiden muuttujien kohdalla puuttuvuus on suurta itse asiassa sen vuoksi, että jos henkilö on vastannut kysymykseen 0, on hänen vastauksensa tallennettu puuttuvaksi. Kaikille tarkoitettujen ei-pakollisten kysymysten puuttuvuus on esitetty kuvassa 5.1.

Taulukosta B huomataan, että kaikissa pakollisissa kysymyksissä puuttuvuus on 0. Tämä onkin loogista, sillä verkkokysymyslomake ei päästä vastaajaa eteenpäin, mikäli kaikkiin pakollisiin kysymyksiin ei ole vastattu. Kaikki lomakkeet täytetään sähköisessä muodossa. Mahdollisuutena on myös täyttää paperinen lomake, mutta tällöinkin tutkimushenkilökuntaan kuuluva henkilö tallentaa vastaukset sähköiseen muotoon. Pakollisiin kysymyksiin on siis pakko vastata tällaisissakin tilanteissa, jolloin näiden kysymysten kohdalla ei puuttuvuutta esiinny.

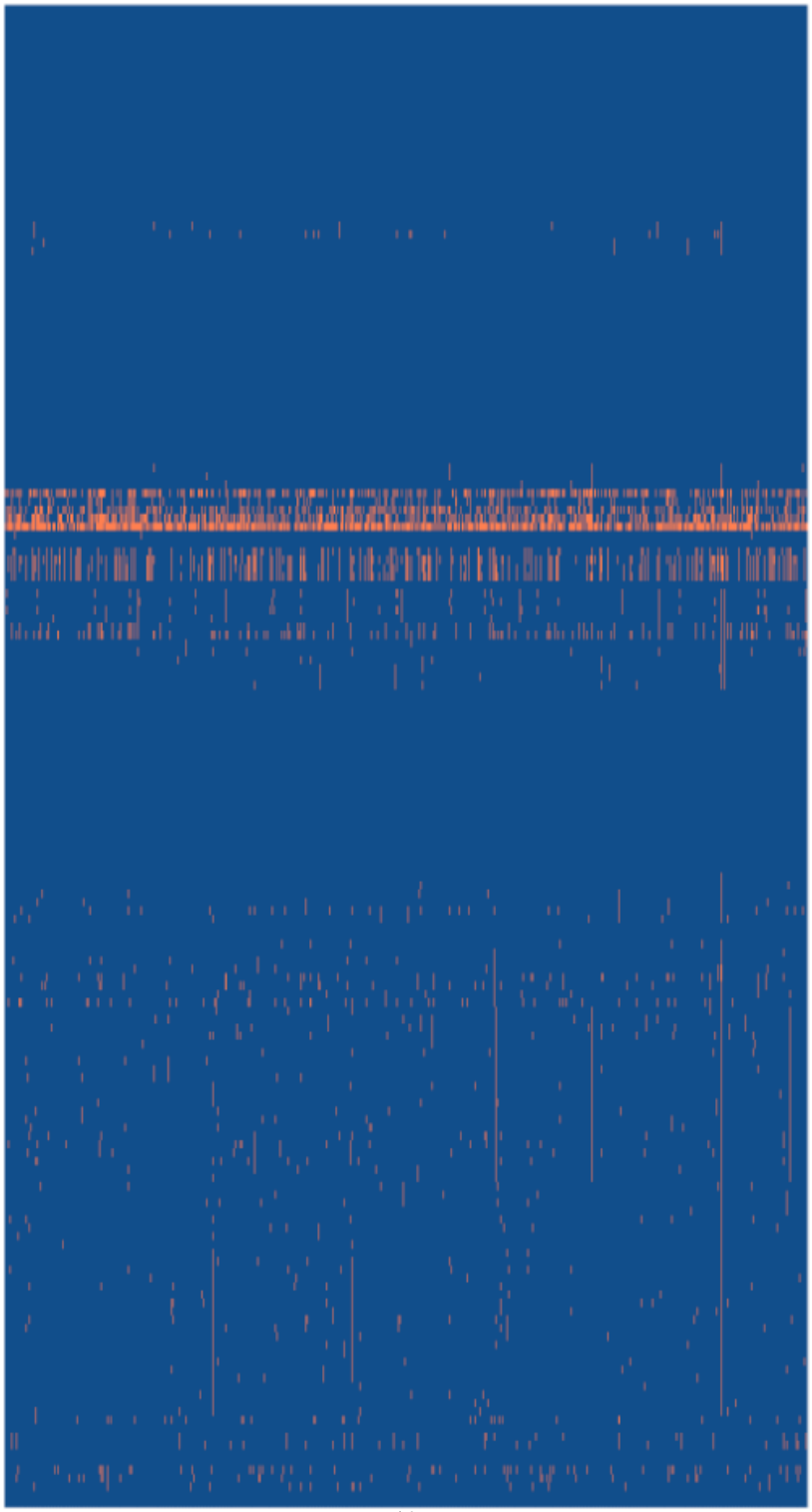
Nähdään, että myös kaikille osoitetuissa kysymyksissä

- 10.4 ”Keneltä olette saanut ammattiapua elintapamuutoksienne tukemiseksi?”
- 10.6 ”Oletteko osallistunut seuraaviin ryhmätoimintoihin ja/tai valmennusohjelmiin viimeksi kuluneen viiden vuoden aikana ainakin jonkin aikaa säännöllisesti? Yksittäistä luentoa, keskustelutilaisuutta, ryhmätuntia tai palveluun rekisteröitymistä ei oteta huomioon”

puuttuvuus on 0. Kysymyslomaketta tarkasteltaessa huomataan, että nämä kaikki kysymykset ovat sellaisia, joista löytyy vastausvaihtoehtoja. Arvot tallennetaan dataan siten, että jos vastausvaihtoehdoista valitsee yhden tai useamman, tallennetaan arvo 1, ja jos ei valitse yhtään, tallennetaan 0.

Aineiston tallentajalla onkin keskeinen rooli myös puuttuvuuden käsittelyn osalta. Ei ole samantekevää, miten puuttuvat arvot koodataan aineistoon, koska kuten tämän aineiston kohdalla huomataan, siinä saatetaan jopa hävittää aineistossa olemassa olevia arvoja. Jälkeen päin voi olla mahdotonta päätellä, onko joku tietty nolla-arvo puuttuva vai todellinen nolla-arvo.

Tarkasteltiin puuttuvuutta vielä puuttuvuutta vastaajakohtaisesti kuvassa 5.2.
[Honaker et al., 2011] Ylhäältä katsottuna ensimmäisen paljon puuttuvuutta sisältävän alueen kysymykset liittyvät istumiseen, seuraavan tupakointiin ja kolmannen uneen. Huomataan, että loppua kohden esiintyy hajanaista puuttuvuutta, ja kysymyslomakkeen alussa ei juuri lainkaan. Muutamat vastaajat ovat jättäneet pitkiä kysymyssarjoja vastaamatta. Puuttuvuuden tarkastelu graafisesti on hyödyllistä, koska jo yhdellä silmäyksellä nähdään, mihin puuttuvuus on keskittynyt, ja onko esimerkiksi jotain selkeää kaavaa, jota se noudattaa.



Kuva 5.2: Puuttuvuuskartta, jossa puuttuvat korallinvärisellä, ja ei-puuttuvat havainnot tummansinisellä. Vaakasuunnassa kysymykset ylhäältä alas ja pystysuunnassa vastaajat. Mukana vain kaikille osoitetut, ei-pakolliset kysymykset.

5.1.1 Mistä puuttuvuus johtuu?

Kysymyksen vastausvaihtoehdot saattavat olla sellaisia, ettei minkään niistä koeta pätevän omalla kohdalla. Kysymyksistä saattaa myös puuttua 'en osaa sanoa' -vaihtoehto, jolloin kysymys jätetään kokonaan välistä, mikäli siihen ei osata vastata.

Kysymyslomake voidaan muotoilla siten, että kysymyksiä ei voi jättää välistä, vaan jokaiseen on vastattava päästäkseen eteenpäin lomakkeella. Toisaalta tällaisen lomakkeen voisi jo itsessään ajatella aiheuttavan sitä, että henkilö päättää jättää lomakkeen täytön kokonaan kesken siinä kohtaa, kun johonkin kysymykseen ei osata vastata tai lomake koetaan esimerkiksi liian pitkäksi. Tutkimuksesta riippuu, tallentuvatko tällaiset kesken jätetyt lomakkeet dataan vai eivät.

Toisinaan saattaa käydä myös niin, että kysymykseen jätetään vastaamatta tai vastataan ei-toivotulla tavalla kysymyksen asettelun epäselvyyden vuoksi. Voi olla, että kysymystä ei ymmärretä tai sen voi ymmärtää monella tavalla. Esimerkiksi tunnetiloihin liittyvät kysymykset voidaan vastaajasta riippuen ymmärtää monin eri tavoin.

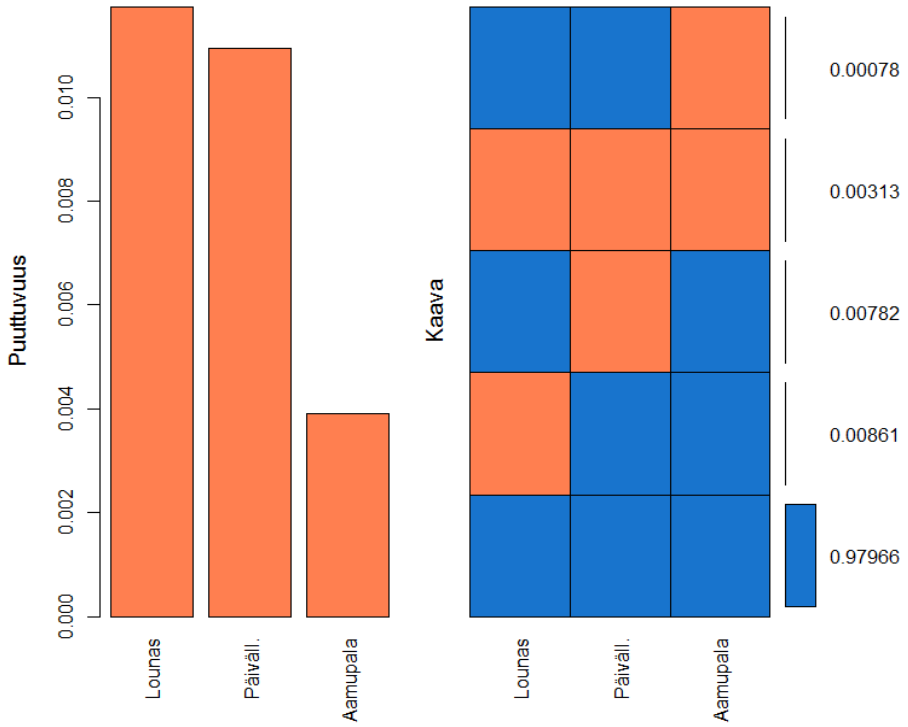
Joskus kysymyslomakkeet sisältävät niin sanottuja hyppykysymyksiä, jossa käsketään tietyn vastausvaihtoehdon valinnan jälkeen hyppäämään yli joitakin kysymyksiä. Lomakkeessa on tällöin ilmaistava yksiselitteisesti, mihin kysymykseen on hypättävä (ellei verkkolomake huolehdi hypystä automaattisesti, kuten tässä lomakkeessa käytettävän lomakkeen tapauksessa). Jos ohjeet ovat epäselvät, saattaa vastaaja esimerkiksi hypätä yli liian monta kysymystä.

Kysymysten asettelusta voi seurata systemaattista virhettä esimerkiksi silloin, kun kysymykset ovat monitulkintaisia. Esimerkiksi tunnetiloihin viittaavat kysymykset voidaan ymmärtää vastaajasta riippuen monin eri tavoin. [Tiirikainen, 2006]

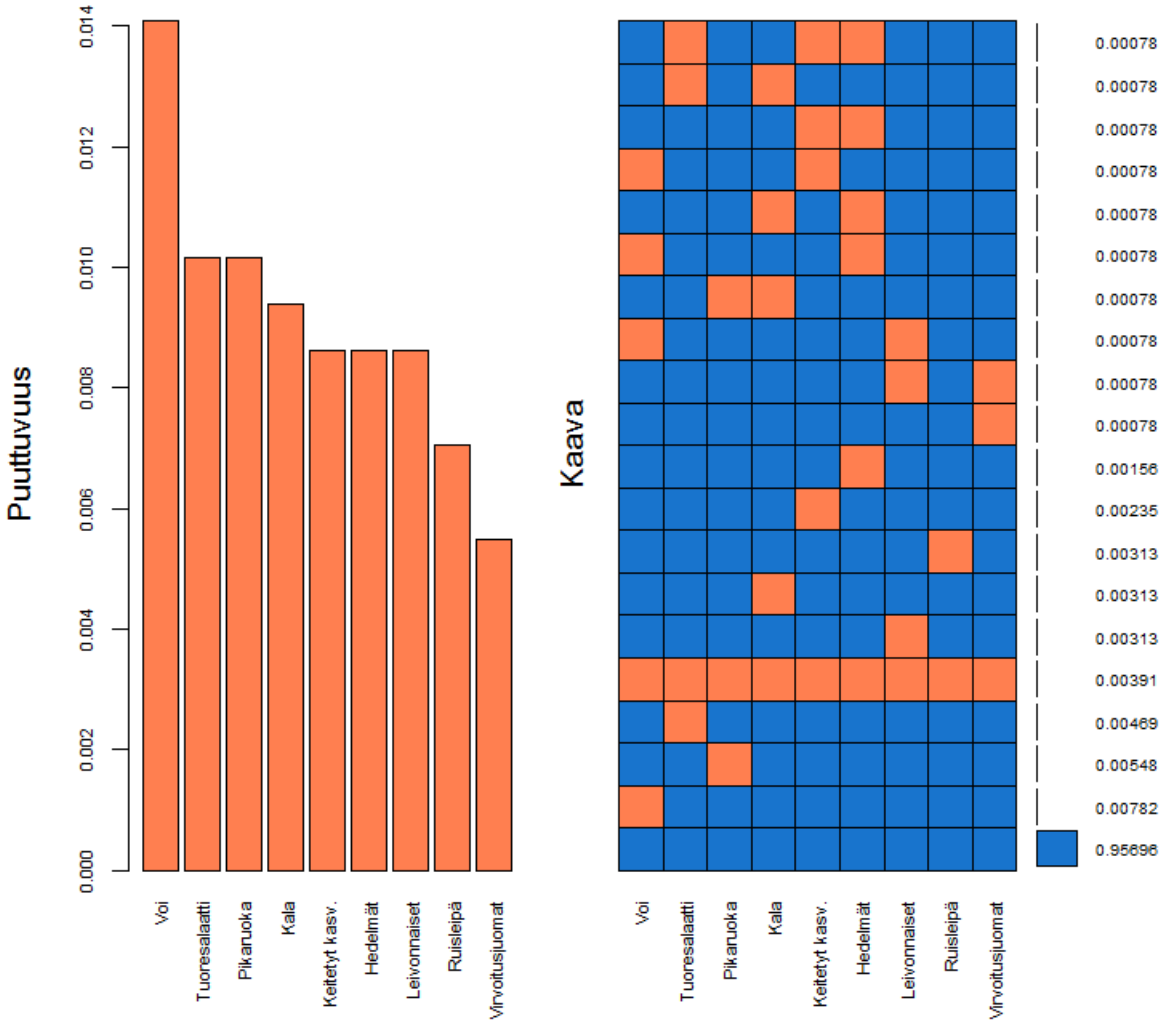
5.2 Puuttuvuus kiinnostuksen kohteena olevissa muuttujissa

Tarkasteltiin puuttuvuutta mielenkiinnon kohteena olevissa kysymyssarjoissa. Mielenkiintoisena voidaan pitää esimerkiksi sitä, millä tavoin vastaaja on jättänyt vastaamatta kysymyksiin: Onko väliin jätetty sarjan kaikki kysymykset, onko johonkin kysymykseen jätetty useammin vastaamatta kuin johonkin toiseen tai esiintyykö kysymysten loppua kohden ns. vastausväsymystä. Yksinkertaisen kuvatarkastelun avulla voidaan saada myös viitteitä esimerkiksin siitä, onko kysymyssarjassa sellaisia kysymyksiä, jotka suurin osa vastaajista on katsonut olevan sellaisia, etteivät ne koske heitä ja jättäneet siksi vastaamatta. Kuvan perusteella voidaan sitten alkaa tutkimaan muuttujassa esiintyvää puuttuvuutta tarkemmin.

Piirrettiin kysymyssarjasta 11.1A-11.1C kuviopari 5.3, josta käy ilmi puuttuvuuden kaava, ts. kuinka monella eri tavalla vastaajat ovat voineet jättää vastaamatta kysymyksiin. [Kowarik and Templ, 2016] Pylväskuvassa nähdään kysymyskohtaisesti suhteellinen osuus vastaamattomuudelle. Esimerkiksi n. 1.2 % vastaajista on jättänyt vastaamatta lounasta koskevaan kysymykseen. Toisessa kuvaajassa nähdään esitetään keltaisella puuttuvia, ja sinisellä vastattuja. Nähdään, että 98.0 % vastaajista on vastannut kaikkiin kolmeen kysymykseen, kun taas 0.3 % ei ole vastannut niistä yhteenkään. Eri ruokalajien syömistä koskevilla kysymyksillä malli ovat hieman monimutkaisempi. Kysymyksiä on enemmän, joten erilaisia vastaamattomuuden kombinaatioitakin on enemmän. Malli on esitetty kuvassa 5.4. Kuvaajia tulkitaan samoin kuin kuvassa 5.3.



Kuva 5.3: Puuttuvuuden kaava, pääateriakysymykset. Kysymys: Kuinka monena arkipäivänä viikossa syötte seuraavat pääateriat?



Kuva 5.4: Puuttuvuuden kaava, ruokalajikysymykset. Kysymys: Kuinka usein tavallisesti käytätte seuraavia elintarvikkeita?

5.3 Tutkimusaineiston imputointi

Kiinnostuksen kohteena olevat muuttujat on esitelty aiemmin alaluvussa 4.4.1. Ne on koottu yhdeksi aineistoksi. Näiden muuttujien lisäksi on aineistoon lisätty toisesta GeneRISK-tutkimukseen liittyvästä lomakeaineistosta tutkittavien painot. Jotta imputointimalleja

voitaisiin verrata todelliseen aineistoon, poistettiin tästä aineistosta kaikki ne rivit, jotka sisältävät yhtään puuttuvuutta. Aineistoon jäi jäljelle 1078 riviä.

Tämän jälkeen aineistoon simuloitiin vuoron perään kaikkia puuttuvuuden tyyppejä. Ensin simuloitiin täysin satunnaista puuttuvuutta (MCAR) poistamalla jokaisesta kiinnostuksen kohteena olevasta muuttujasta satunnaisesti 0-500 havaintoa.

Satunnaista puuttuvuutta (MAR) simuloitiin poistamalla havaintoja kullekin muuttujalle yksilöidyn kaavan mukaan. Puuttuvuus ei riippunut puuttuvuutta sisältävästä muuttujasta itsestään, vaan muista mallissa mukana olevista muuttujista. Esimerkiksi kysymykselle ”Kuinka paljon liikut ja rasitat itseäsi ruumiillisesti vapaa-aikana” havainnon puuttumisen todennäköisyys oli 0.9 jos vastaaja on yli 60-vuotias ja 0.4 muuten. Tässä siis tehtiin oletus, että yli 60-vuotiaat eivät haluaisi vastata kysymykseen yhtä mielellään kuin nuoremmat.

Ei-satunnaista puuttuvuutta (MNAR) simuloitiin asettamalla puuttuvuus riippuvaiseksi muuttujasta itsestään. Siis esimerkiksi kysymykselle ”Kuinka paljon liikut ja rasitat itseäsi ruumiillisesti vapaa-aikana” havainnon puuttumisen todennäköisyys oli 0.9, jos vastaaja oli valinnut pienimmän mahdollisen liikuntamäärän ja 0.2 muutoin. Tässä siis oletettiin, että vähemmän liikkuvat jättävät kysymykseen mieluummin vastaamatta kuin enemmän liikkuvat.

Näin muodostuneiden, puuttuvuutta sisältävien aineistojen muuttujakohtaiset havaintomäärät on esitelty taulukossa 5.1. Seuraavaksi jokaiseen aineistoon sovellettiin aiemmin esiteltyjä imputointimenetelmiä. Imputoinnin onnistumista tarkasteltiin muodostamalla lineaarinen regressiomalli, jossa painoa selitettiin iällä, sukupuolella, liikuntatottumuksilla ja istumisen päivittäisellä määrällä. Lineaarinen regressiomalli muodostettiin aluksi täydelliselle aineistolle ja imputoiduille aineistoille muodostettuja regressiomalleja verrattiin tähän. Tässä tarkastelussa imputointi voidaan katsoa hyvin onnistuneeksi, jos regressiomalli säilyy samankaltaisena: samat selittäjät säilyvät merkitsevinä, kuin täydelliselle

datalle muodostetussa mallissa, selitysaste säilyy samankaltaisena ja beta-kertoimet samankaltaisina.

Toisena tarkasteluna käytettiin yksinkertaista frekvenssijakaumien tarkastelua pylväskuvien avulla.

Taulukko 5.1: Puuttuvien havaintojen määrä kiinnostuksen kohteena olevissa muuttujissa, kun eri tyyppisiä puuttuvuuksia on simuloitu aineistoon. N=1078

Muuttuja	MCAR	MAR	MNAR
Ruokakunnan kokonaistulot	633	591	492
Liikuntatottumukset	528	563	289
Istuminen	526	697	117
Aamiainen	530	629	307
Lounas	540	635	337
Päivällinen	540	442	380
Ruisleipä	554	324	265
Tuorekasvikset	527	339	231
Keitetyt kasvikset	529	332	334
Hedelmät	634	349	168
Kala	525	624	0
Pikaruoat	518	352	213
Voi	536	331	478
Jälkiruoat	511	373	371
Virvoitusjuomat	523	357	240

5.3.1 Ehdollinen moni-imputointi

Ehdolliseen moni-imputointiin käytettiin R:n [R Core Team, 2017] mice-pakettia [van Buuren and Groothuis-Oudshoorn, 2011]. Se toimii siten, että funktioon syötetään puuttuvuutta sisältävä aineisto, ja jokaiselle aineiston muuttujalle valitaan sopiva impu-

tointimenetelmä muuttujan tyypin mukaan: Predictive mean matching jatkuville muuttujille, logistinen regressioimputointi kaksiluokkaisille muuttujille, moniarvoregressioimputointi kategorisille muuttujille ja proportional odds model järjestysasteikollisille muuttujille. Muodostettavien imputoitujen aineistojen määräksi asetettiin 5.

Ehdollisen moni-imputointimallin oletettiin toimivan hyvin sekä täysin satunnaisen puuttuvuuden, että satunnaisen puuttuvuuden tapauksissa. Ei-satunnaisen puuttuvuuden suhteen mallin ei oletettu toimivan kovin hyvin, koska imputointimenetelmälle ei annettu mitään lisätietoa puuttuvuuden mekanismista, vaikka se olikin tässä tapauksessa tiedossa.

5.3.2 JM-moni-imputointi

JM-moni-imputointiin käytettiin R:n Amelia-pakettia [Honaker et al., 2011]. Puuttuvuutta sisältävä aineisto syötetään amelia-funktioon ja aineistosta määritellään, mitkä muuttujat tulisi imputoida kategorisina (loput jatkuvina). Tässä valittiin ehdollisen moni-imputoinnin tapaan muodostettavien imputoitavien aineistojen määräksi 5.

Kirjallisuuden perusteella ehdollisen moni-imputoinnin ja JM-moni-imputoinnin oletettiin toimivan samantasoisesti keskenään. JM-moni-imputoinnin etuna huomattiin tätä tutkielmaa kirjoittaessa, että se toimii huomattavasti nopeammin kuin ehdollinen moni-imputointi. Ehdollinen moni-imputointi R-ohjelmistolla voi kestää useita minuutteja tietokoneen tehosta riippuen, kun taas JM-moni-imputointiin kuluu aikaa vain muutamia sekunteja.

5.3.3 Keskilukuimputointi

Keskilukuimputoinnissa kaikkien imputointiaineistoon valittujen puuttuvuutta sisältävien muuttujien puuttuvien arvojen paikalle sijoitettiin muuttujien mediaanit.

Tämän voitaisiin olettaa toimivan parhaiten silloin, kun puuttuvuus on täysin satunnaista, jolloin puuttuvuus ei vääristä muuttujien välisiä keskinäisiä suhteita eikä aiheuta

aineistoon harhaa. Tällöin imputoinnin tarkoitus on täydentää aineistoa siten, että lineaarisessa regressioanalyysissa kaikki rivit saadaan mukaan, menetelmä kun jättää ulkopuolelle kaikki puuttuvuutta sisältävät rivit.

5.3.4 CC-analyysi

Kun imputoiduista aineistoista poistettiin kaikki ne rivit, jotka sisälsivät puuttuvuutta, jäljelle ei jäänyt yhtään riviä täysin satunnaisen puuttuvuuden tapauksessa, yksi satunnaisen puuttuvuuden tapauksessa ja kuusi ei-satunnaisen puuttuvuuden tapauksessa. Jo tämä kertoo siitä, kuinka radikaalisti käytettävissä olevien rivien määrä voi pudota, jos analyyseissa hyödynnetään ainoastaan täydellisesti havaittuja rivejä.

Koska lineaarisessa regressiomallissa käytettiin vain muutamaa muuttujaa, voitiin aineistosta hyödyntää kuitenkin useampia täydellisiä rivejä. CC-aineiston voitaisiin olettaa toimivan huonosti verrattuna imputoituihin aineistoihin, koska havaintoja on menetetty niin paljon.

5.3.5 Looginen imputointi

Jotta myös loogisen imputoinnin idea tulisi havainnollistettua, käsitellään sitä seuraavaksi lyhyen esimerkin avulla. Tätä imputointimallia ei kuitenkaan oteta mukaan vertailuihin, koska se sopii käytettäväksi vain tietynlaisiin muuttujiin. Loogista imputointia voitaisiin hyödyntää vaikkapa muuttujien

- 2.3 ”Kuinka monta kokonaista päivää olitte viimeksi kuluneen vuoden (12 kk) aikana sairauden takia poissa töistä tai hoitamatta tavallisia tehtäviänne? (Jos ette yhtään, vastatkaa 0.)”
- 15.10 ”Kuinka moni taloutenne jäsenistä on 0-13v?”

”paikkaamiseksi”. Voidaan tehdä karkea oletus siitä, että sellaisten henkilöiden kohdalla, jotka ovat jättäneet vastaamatta näihin kysymyksiin, vastaus kysymyksiin olisi 0. Tämä on kuitenkin vain oletus ja riippuu muuttujien jatkoanalyysitarkoituksista, kannattaako

sitä tehdä.

Jos verrataan nollalla imputoidun, talouden 0-13-vuotiaiden määrää koskevan, muuttujan tunnuslukuja alkuperäisen muuttujan tunnuslukuihin, voidaan niissä huomata suuria eroja:

Taulukko 5.2: Alkuperäisen ja imputoidun muuttujan tunnusluvut.

	Minimi	1.kvartiili	Mediaani	Keskiarvo	3.kvartiili	Maksimi	Puuttuvat
Alkuperäinen	1.000	1.000	1.000	1.422	2.000	4.000	1208
Imputoitu	0.000	0.000	0.000	0.195	0.000	4.000	0

Jos todellinen tilanne on oletusten vastainen, muuttujan jakauma vääristyy huomattavasti imputoinnin seurauksena. On siis syytä pitää mielessä mahdollisuus, että myös osa heistä, joiden vastaus kysymykseen on jotain muuta kuin 0, on vain jättänyt syystä tai toisesta vastaamatta kysymykseen.

5.3.6 Tulokset

Lineaaristen mallien vertailu tuotti yllättäviä tuloksia. Tarkoituksena oli osoittaa, että moni-imputointi toimisi paremmin kuin muut menetelmät, mutta tutkielmassa käytetyn aineiston perusteella tällaista johtopäätöstä ei voida vetää. Kirjallisuuden perusteella voitiin todeta, että jos puuttuvuus on täysin satunnaista, ei ole juurikaan väliä sillä, poistetaanko puuttuvuutta sisältävät havaintorivit aineistosta kokonaan, vai imputoidaanko puuttuvien arvojen paikalle arvot jollain tässäkin tutkielmassa mainituista imputointimenetelmistä. Tämä seikka pyrittiin ottamaan huomioon simuloimalla erilaisia puuttuvuuksia.

Täysin satunnaisen puuttuvuuden tapauksessa (5.4) tulokset olivatkin odotusten mukaisia: Lineaariset regressiomallit tuottivat keskenään samankaltaisia kertoimia ja selitysas-

teita. Mediaani-imputointi ja complete-case antoivat tosin täysin satunnaisen ja satunnaisen (5.5) puuttuvuuden tilanteissa liikunnalle suuremman negatiivisen kertoimen kuin moni-imputointimenetelmät. Kun kertoimia verrattiin alkuperäisestä aineistosta lasketuun regressiokertoimeen, huomattiin, että mediaani-imputointi ja complete case -analyysi onnistuivat itse asiassa estimoimaan liikunnan kerrointa paremmin kuin moni-imputointi. Lisäksi voitiin huomata, että täydellisesti satunnaisen puuttuvuuden tapauksessa moni-imputointimenetelmät nostivat virheellisesti myös iän merkitseväksi painoa selittäväksi tekijäksi, vaikkeivät cc-imputointi tai mediaani-imputointi tätä tee.

Täysin satunnaisen puuttuvuuden kohdalla ei pitäisi olla merkityksellistä, käytetäänkö yksinkertaista vai moni-imputointia. Rivit voitaisiin jättää jopa kokonaan pois analyyseistä, tosin analyysien voima voi heiketä havaintojen vähentyessä. Sen sijaan satunnaisen puuttuvuuden ja ei-satunnaisen puuttuvuuden tilanteissa on kirjallisuuden perusteella tärkeää, minkä imputointimenetelmän valitsee. Tässä tutkielmassa ei kuitenkaan odotetun kaltaisia eroja havaittu. Verrattaessa lineaarisia regressiomalleja kaikkien kolmen puuttuvuuden tyypin tilanteissa, huomataan, ettei niiden välillä ole juurikaan eroja. Yllä todetaankin, miten itse asiassa moni-imputointimenetelmät tuottavat jopa huonompia estimaatteja.

Kun tuloksia lähdettiin tutkimaan tarkemmin, huomattiin, että imputoitavien muuttujien välillä ei vallitse juurikaan tilastollisesti merkitseviä yhteyksiä. Tästä johtunee se, että satunnaisen puuttuvuuden simulointi ei muuttanut muuttujien jakaumia niin, että imputointimenetelmien erot olisivat tulleet esille. Esimerkiksi tutkimukseen osallistuneiden yli 60-vuotiaiden liikuntatottumusten jakauma ei poikennut alle 60-vuotiaiden jakaumasta, mikä selittää sen, ettei jakauma muutu, vaikka simuloitaisiin puuttuvuus 90 prosentin todennäköisyydellä yli 60-vuotiaille vastaajille ja 40 prosentin todennäköisyys 60-vuotiaille ja sitä nuoremmille liikuntatottumuksia koskevassa kyselyssä.

Tämä ei kuitenkaan vielä selitä sitä, miksi myös ei-satunnaisen puuttuvuuden tilanteessa (5.6) menetelmät toimivat keskenään yhtä hyvin. Ei-satunnaisen puuttuvuuden tapauk-

sessa puuttuvuutta pyrittiin simuloimaan poistamalla ääriarvoja muuttujista. Näin toimittiin siksi, että näin oletettiin tapahtuvan helposti myös tosielämässä: Esimerkiksi kaikkein vähiten liikkuvat voisivat jättää muita herkemmin vastaamatta liikuntakysymyksiin ja eniten pikaruokaa syövät voisivat jättää muita herkemmin vastaamatta kysymykseen siitä, kuinka usein syövät pikaruokaa. Tämän seurauksena jäljelle jäi paljon keskimääräisiä vastauksia ja vain vähän ääriarvovastauksia. Alkuperäisessä aineistossa suurimmassa osassa muuttujia ääriarvovastauksia oli vain vähän, ja siksi niiden poistamisella ei ollut näkyvää merkitystä lineaarisiin regressioihin. Tästä voidaan päätellä, että näiden kysymysten osalta aineiston vastaajat ovat ns. melko keskimääräisiä, ei ole esimerkiksi niin, että mukaan on valikoitunut vain erityisen liikunnallisia tai erityisen paljon virvoitusjuomia nauttivia henkilöitä.

Kun tarkastellaan pylväskuvia, voidaan erityisesti ruisleivän (5.7) ja tuorekasvisten 5.8) kohdalla tehdä huomio siitä, kuinka JM-imputointi vaikuttaa muokkaavan muuttujien jakaumaa hieman enemmän kohti normaalijakaumaa, kun taas FCS-imputointi ei tätä tee. FCS-imputointi sen sijaan muistuttaa kaikkien pylväskuvien kohdalla hyvin paljon CC-tilannetta. Se siis näyttäisi mukailevan käytettävissä olevaa aineistoa, kun taas JM-imputoinnin menetelmät lisäävät aineistoon enemmän vaihtelua.

Mediaani-imputoinnin huomaa helposti jakaumissa näkyvistä ns. piikeistä mediaanin kohdalla. Tässä voidaan nähdä, kuinka muuttujien jakaumat vääristyvät mediaani-imputointia käytettäessä. Silmämääräisesti voidaankin päätellä, että menetelmä häviää reilusti muille menetelmille jakauman estimoimisessa.

Kaiken kaikkiaan voidaan todeta, että tässä tutkielmassa käytettyjen muuttujien suhteen ei aineistossa vaikuttaisi olevan suuria puuttuvan tiedon ongelmia. Kuten todellisen aineiston kattavassa puuttuvuustarkastelussa huomattiin, puuttuvuus oli näissä muuttujissa vähäistä. Lisäksi vähäiset ääriarvovastaukset, sekä muuttujien keskinäiset riippumattomuudet johtavat siihen, etteivät yksittäiset puuttuvat havainnot tule vaikuttamaan suuresti analyyseihin. Toki puuttuva tieto on aina olemassa olematonta tietoa, ja meillä

on käytettävissä vain tehdyt havainnot, joten on olemassa mahdollisuus, että vähäisetkin puuttuvat havainnot ovat jollain tavalla hyvin systemaattisia. Aineiston on tarkoitus edustaa alueen 45-65-vuotiasta väestöä, mutta siinä ei ole mukana esimerkiksi yhtään vakituisessa sairaalahoidossa olevaa. Heidän vastauksensa joihinkin kysymyksiin saattaisivat olla hyvinkin erilaisia verrattuna valtaväestöön, joka asuu kotona. Näitä on hyvä miettiä kun mietitään imputoinnin tarpeellisuutta lopullisen aineiston kohdalla.

Lineaarinen regressio. Alkuperäinen aineisto

	Riippuva muuttuja:
	paino
Ikä	-0.086 (0.077)
Sukupuoli: mies	16.104^{***} (0.890)
Liikuntatottumukset	-3.658^{***} (0.458)
Istuminen	0.009^{***} (0.002)
Vakiotermi	84.949^{***} (4.745)
Havaintojen määrä	1 078
R^2	0.289
Adjustoitu R^2	0.286
Jäännösvirhe	14.360 (df = 1073)
F Statistiikka	108.831^{***} (df = 4; 1073)
Huomiot:	*p<0.1; **p<0.05; ***p<0.01

Taulukko 5.3: Lineaarinen malli. Alkuperäinen aineisto

Lineaarinen regressio. MCAR-puuttuvuus

	Riippuva muuttuja:			
	paino			
	FCS	CC	JM	mediaani
	(1)	(2)	(3)	(4)
Ikä	-0.061** (0.034)	-0.059 (0.145)	-0.067** (0.035)	-0.102 (0.078)
Sukupuoli: mies	15.878*** (0.395)	15.258*** (1.632)	16.110*** (0.399)	16.306*** (0.901)
Liikuntatottumukset	-3.199*** (0.194)	-4.284*** (0.831)	-2.531*** (0.176)	-4.103*** (0.650)
Istuminen	0.014*** (0.001)	0.014** (0.004)	0.012*** (0.001)	0.011** (0.003)
Vakiotermi	80.194*** (2.120)	83.390*** (9.189)	79.506*** (2.113)	86.733*** (5.141)
Havaintojen määrä	5 390	297	5 390	1 078
R^2	0.297	0.315	0.280	0.267
Adjustoitu R^2	0.296	0.305	0.279	0.264
Jäännösvirhe	14.253 (df = 5385)	13.585 (df = 292)	14.424 (df = 5385)	14.581 (df = 1073)
F Statistiikka	567.607*** (df = 4; 5385)	33.492*** (df = 4; 292)	522.512*** (df = 4; 5385)	97.509*** (df = 4; 1073)
Huomiot:				*p<0.1; **p<0.05; ***p<0.01

Taulukko 5.4: Lineaarinen malli. MCAR-puuttuvuus

Lineaarinen regressio. MAR-puuttuvuus

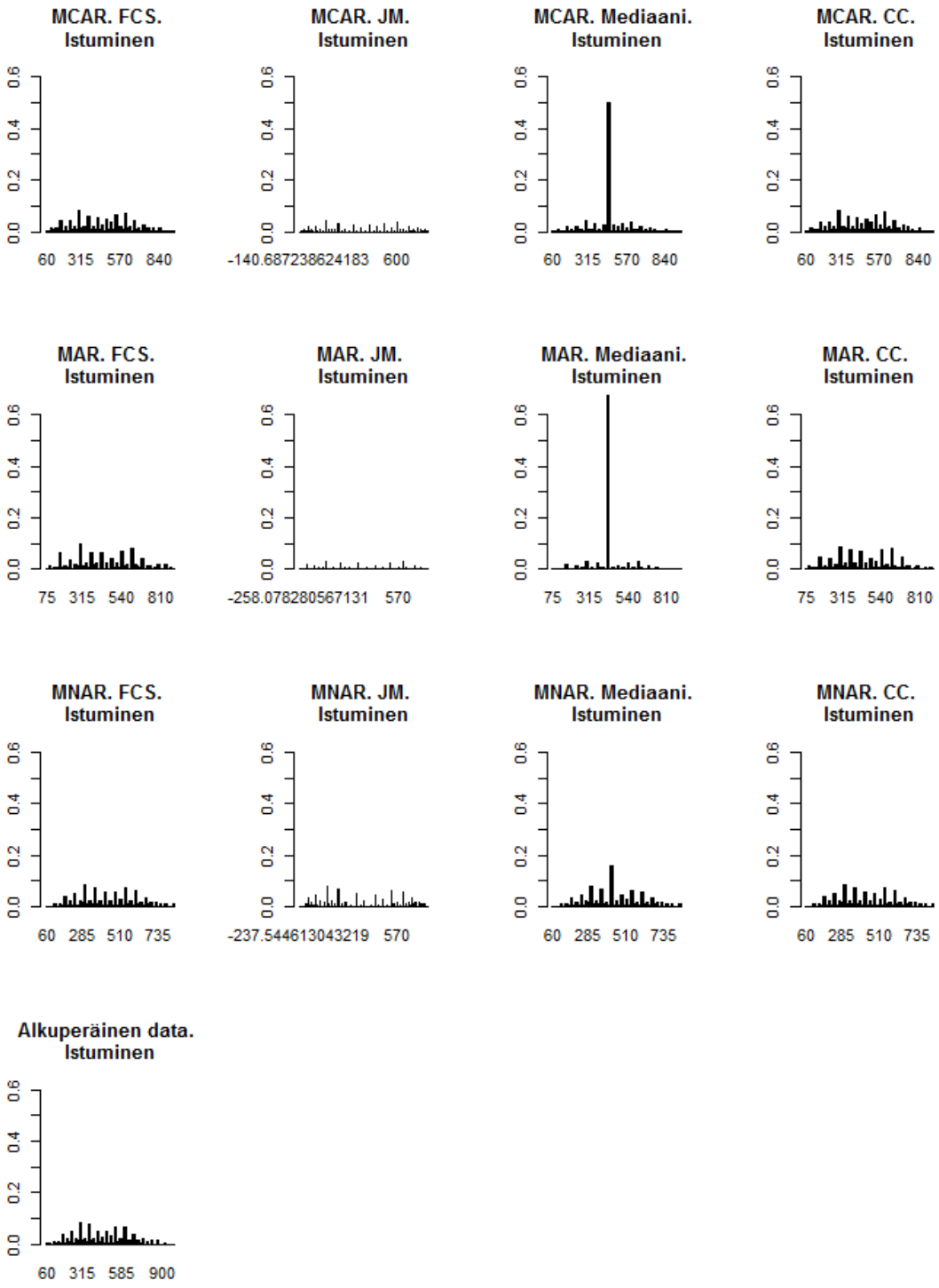
	Riippuva muuttuja:			
	paino			
	FCS	CC	JM	mediaani
	(1)	(2)	(3)	(4)
Ikä	-0.070** (0.034)	-0.079 (0.216)	-0.078** (0.034)	-0.096 (0.078)
Sukupuoli: mies	16.429*** (0.400)	16.258*** (2.143)	16.374*** (0.402)	16.408*** (0.908)
Liikuntatottumukset	-2.695*** (0.191)	-3.337*** (1.205)	-2.502*** (0.175)	-3.349*** (0.690)
Istuminen	0.010*** (0.001)	0.009 (0.006)	0.010*** (0.001)	0.012*** (0.004)
Vakiotermi	81.105*** (2.059)	84.485*** (12.920)	80.757*** (2.079)	83.972*** (5.143)
Havaintojen määrä	5 390	179	5 390	1 078
R^2	0.275	0.282	0.274	0.253
Adjustoitu R^2	0.274	0.266	0.273	0.251
Jäännösvirhe	14.475 (df = 5385)	13.964 (df = 174)	14.483 (df = 5385)	14.712 (df = 1073)
F Statistiikka	509.450*** (df = 4; 5385)	17.116*** (df = 4; 174)	507.215*** (df = 4; 5385)	91.005*** (df = 4; 1073)
Huomiot:				*p<0.1; **p<0.05; ***p<0.01

Taulukko 5.5: Lineaarinen malli. MAR-puuttuvuus

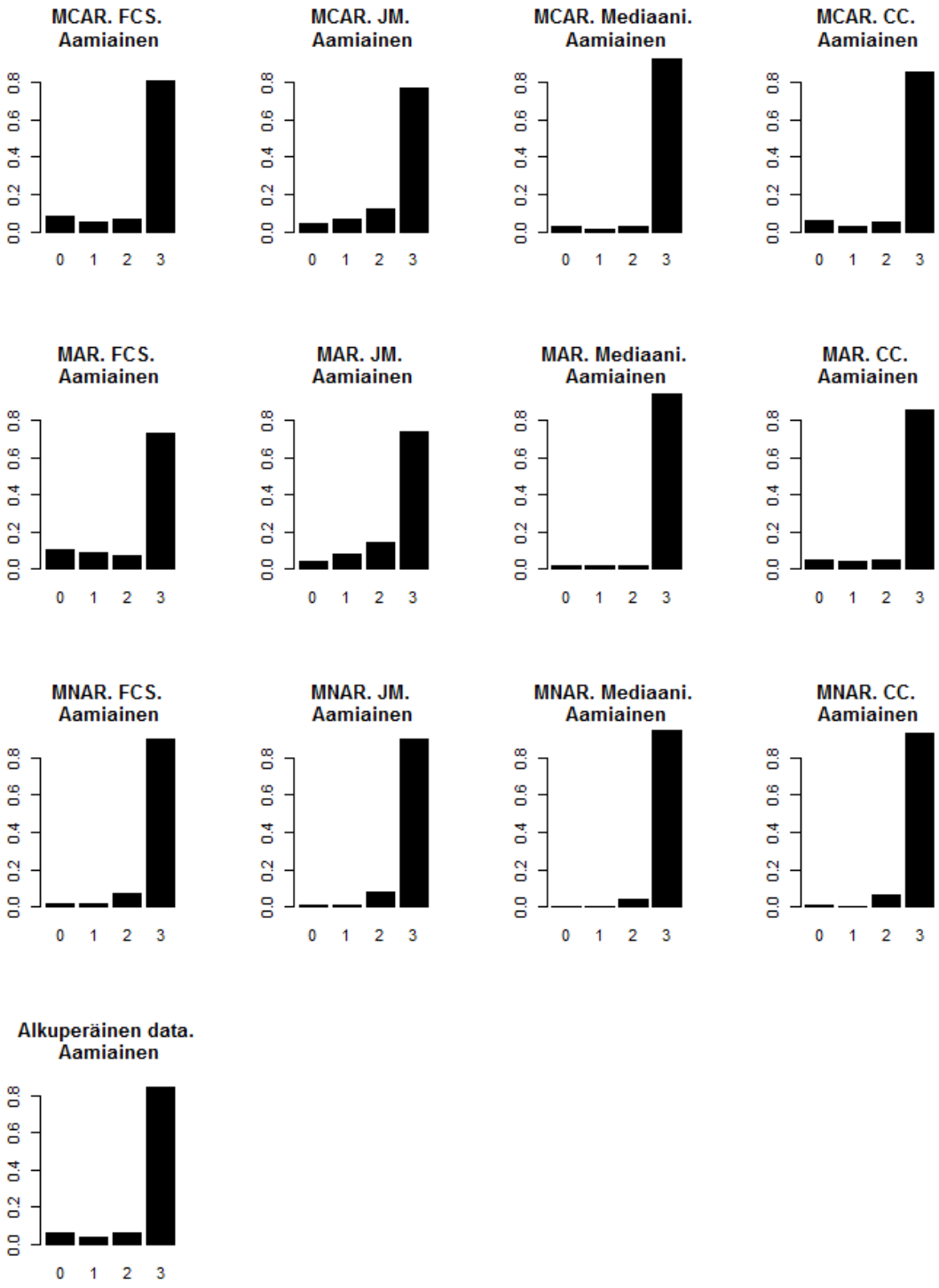
Lineaarinen regressio. MNAR-puuttuvuus

	Riippuva muuttuja:			
	paino			
	FCS	CC	JM	mediaani
	(1)	(2)	(3)	(4)
Ikä	-0.081** (0.034)	-0.080 (0.094)	-0.072** (0.035)	-0.081 (0.078)
Sukupuoli: mies	16.106*** (0.401)	15.197*** (1.087)	16.122*** (0.403)	16.187*** (0.905)
Liikuntatottumukset	-3.760*** (0.238)	-3.353*** (0.686)	-3.059*** (0.216)	-3.791*** (0.659)
Istuminen	0.010*** (0.001)	0.008** (0.003)	0.010*** (0.001)	0.011** (0.003)
Vakiotermi	85.338*** (2.164)	84.539*** (5.924)	82.675*** (2.151)	85.081*** (5.027)
Havaintojen määrä	5 390	700	5 390	1 078
R^2	0.278	0.254	0.273	0.267
Adjustoitu R^2	0.277	0.250	0.272	0.264
Jäännösvirhe	14.440 (df = 5385)	14.111 (df = 695)	14.494 (df = 5385)	14.580 (df = 1073)
F Statistiikka	518.365*** (df = 4; 5385)	59.273*** (df = 4; 695)	504.587*** (df = 4; 5385)	59.273*** (df = 4; 1073)
Huomiot:				*p<0.1; **p<0.05; ***p<0.01

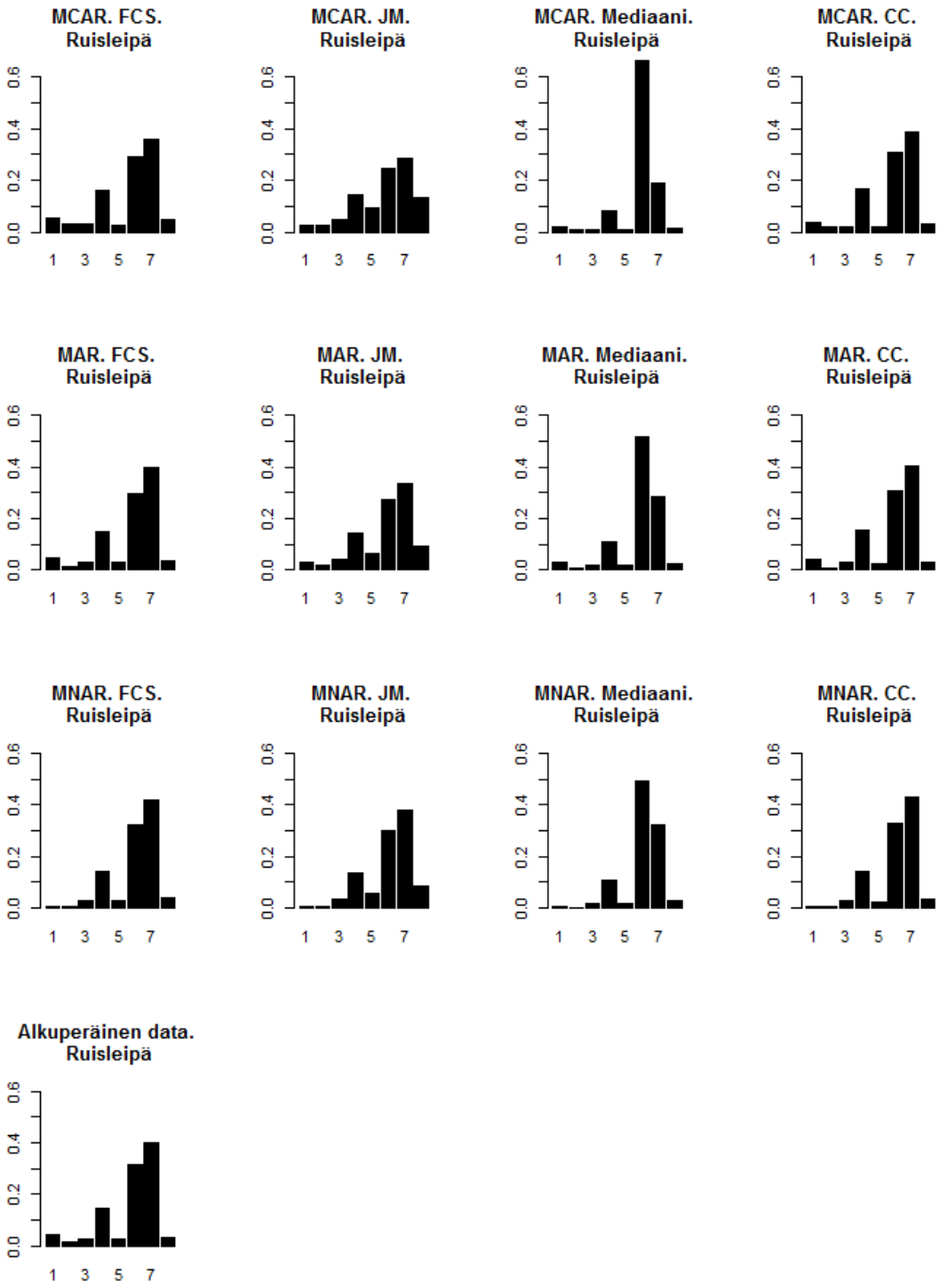
Taulukko 5.6: Lineaarinen malli. MNAR-puuttuvuus



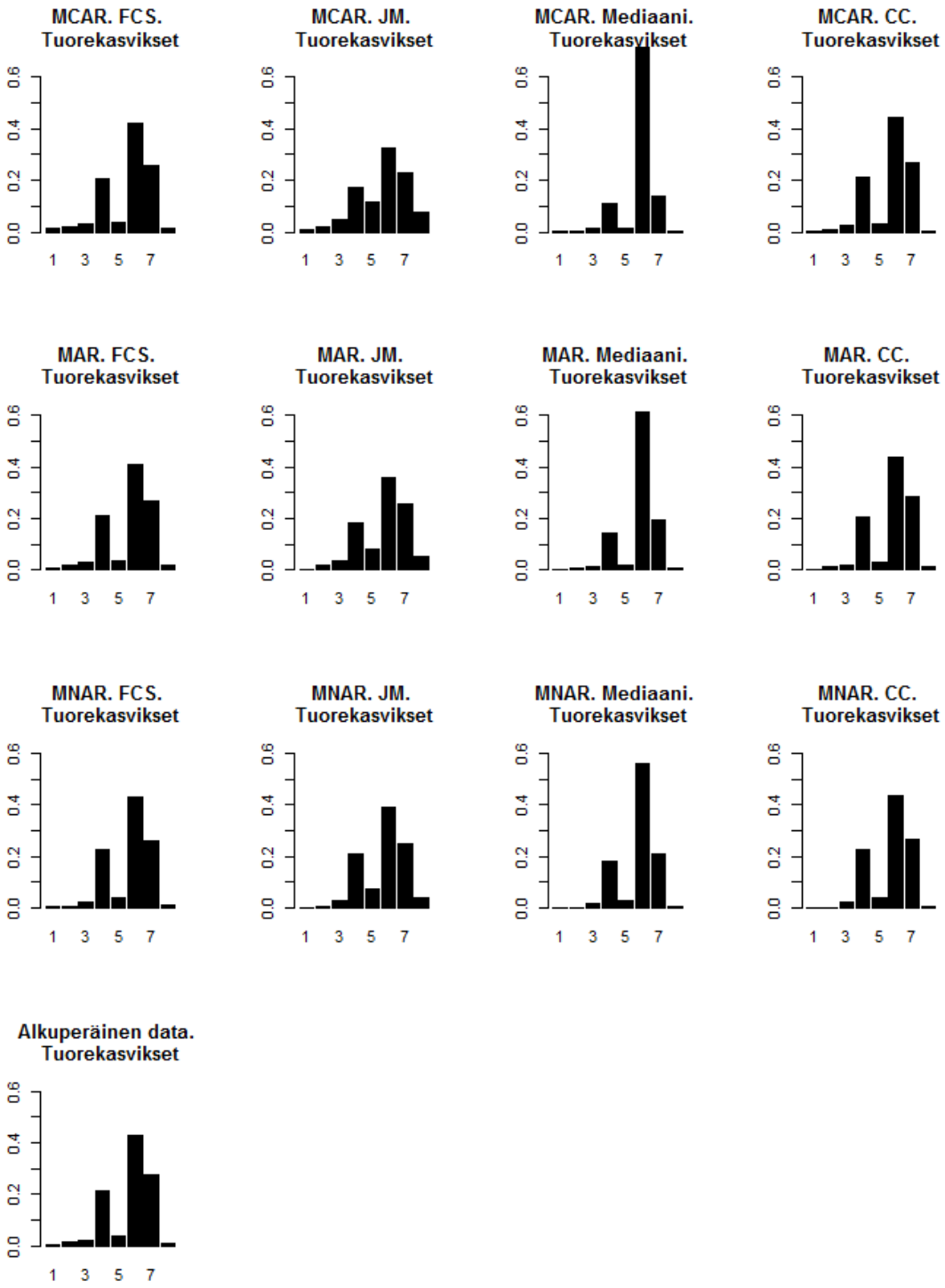
Kuva 5.5: Pylväskuvat, istuminen. Pystyakselilla vastausvaihtoehdon valinneiden suhteellinen osuus ja vaaka-akselilla minuutit.



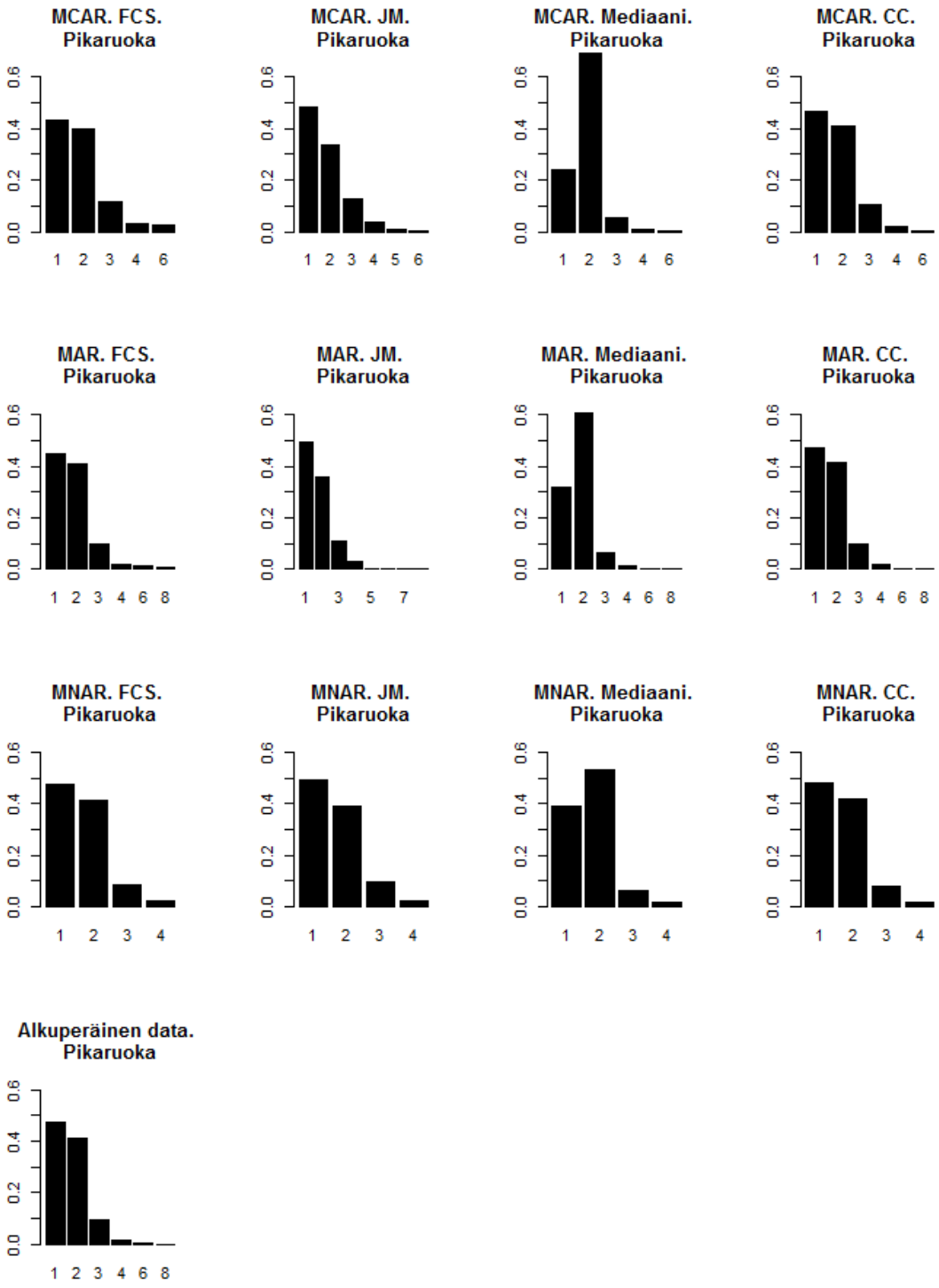
Kuva 5.6: Pylväskuvat, aamiainen. Pystyakselilla vastausvaihtoehdon valinneiden suhteellinen osuus ja vaaka-akselilla vastausvaihtoehdot.



Kuva 5.7: Pylväskuvat, ruisleipä. Pystyakselilla vastausvaihtoehdon valinneiden suhteellinen osuus ja vaaka-akselilla vastausvaihtoehdot.



Kuva 5.8: Pylväskuvat, tuorekasvikset. Pystyakselilla vastausvaihtoehdon valinneiden suhteellinen osuus ja vaaka-akselilla vastausvaihtoehdot.



Kuva 5.9: Pylväskuvat, pikaruoka. Pystyakselilla vastausvaihtoehdon valinneiden suhteellinen osuus ja vaaka-akselilla vastausvaihtoehdot.

Luku 6

Yhteenveto

Menetelmiä puuttuvan datan paikkaamiseksi on useita. Kategorisen datan imputoimista on kuitenkin tutkimuksissa käsitelty vähemmän kuin jatkuvan, ja erityisesti normaalisti jakautuneen datan. Myös menetelmiä kategorisen datan paikkaamiseksi on vähemmän.

Parhaaksi todetuissa imputointimenetelmissä on keskeistä löytää muuttujia, joilla selittää puuttuvuutta sisältävän muuttujan vaihtelua. Tässä tutkimusaineistossa ongelmalliseksi osoittautui se, että aineistossa oli vain vähän sellaisia selittäjiksi kelpaavia muuttujia, jotka eivät sisältäisi puuttuvuutta. Jotta imputointia voitiin selkällä tavalla esitellä aineiston avulla, siitä poistettiin kaikki sellaiset rivit, joilla taustamuuttujat tai osa niistä puuttuivat. Tämä ei kuitenkaan ole todellisessa tilanteessa mahdollinen ratkaisu, koska imputoinnilla pyritään nimenomaan lisäämään käyttökelpoisten rivien määrää, ja tällä menetelmällä menetettiin huomattava määrä rivejä.

Jatkoa ajatellen olisi suositeltavaa päästä käsiksi esimerkiksi tutkittavien rekisteritietoihin, jotta voitaisiin täydentää väestöryhmään liittyviä muuttujia ja käyttää tämän jälkeen näitä selittäjinä. Tällöin päästäisiin siihen todelliseen tilanteeseen, jolloin taustamuuttujat eivät sisältäisi puuttuvuutta ja imputointi voidaisiin suorittaa näitä hyödyntämällä.

Imputointimenetelmistä tehokkaimmaksi osoittautui kirjallisuuskatsauksen peruseteella

moni-imputointi, koska se ottaa huomioon myös epävarmuuden imputoitaessa, toisin kuin yksinkertaiset imputointimenetelmät. Samaan viittaavia tuloksia ei kuitenkaan saatu tutkielmassa käytetystä aineistosta suurelta osin siksi, että imputoitavat muuttujat olivat liian riippumattomia toisistaan, sekä siksi, että aineisto koostui keskenään melko samankaltaisista vastaajista.

Johtopäätöksenä voidaankin todeta, että käytettävällä imputointimenetelmällä on merkitystä erityisesti silloin, kun aineiston muuttujat riippuvat toisistaan ja puuttuvuus riippuu aineiston muuttujista. Tässä tutkielmassa tehtiin vain suppea katsaus aineiston muuttujiin imputointimielessä, joten jatkotutkimuksena voitaisiin tarkastella myös muita aineiston muuttujia.

Tämän tutkielman perusteella ei voida missään tapauksessa kumota kirjallisuudessa aiemmin todettua moni-imputoinnin ylivoimaisuutta, vaan voidaan todeta, että sama ei käynyt ilmi tässä nimenomaisessa aineistossa. Jos lähdettäisiin tutkimaan aineiston muita muuttujia, tai suoritettaisiin samantyyppisiä analyysejä eri aineistoille, voitaisiin päätyä enemmän kirjallisuuden kanssa yhteneviin tuloksiin. Tulokset-luvussa onkin etsitty syitä sille, miksi tämä tutkielma ei päätynyt samankaltaisiin tuloksiin.

Viitteet

[Allison, 2009] Allison, P. (2009). Why you probably need more imputations than you think. http://statisticalhorizons.com/more-imputations. Viitattu: 18.7.2016.

[Andridge, 2010] Andridge, Rebecca R. & Little, R. J. A. (2010). A review of hot deck imputation for survey non-response. *International statistical review = Revue internationale de statistique*, 78(1):40–64.

[Gelman et al., 1995] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis (Chapman & Hall/CRC Texts in Statistical Science)*. Chapman and Hall/CRC.

[Gelman and Hill, 2007] Gelman, A. and Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*, volume Analytical methods for social research. Cambridge University Press, New York.

[GeneRISK-verkkosivut, a] GeneRISK-verkkosivut. Ajankohtaista. http://www.generisk.fi/node/ajankohtaista. Viitattu: 9.3.2016.

[GeneRISK-verkkosivut, b] GeneRISK-verkkosivut. Asiantuntijalle. http://www.generisk.fi/content/asiantuntijalle. Viitattu: 9.3.2016.

[GeneRISK-verkkosivut, c] GeneRISK-verkkosivut. Generisk-tutkimuksesta. http://www.generisk.fi/content/generisk-tutkimuksesta. Viitattu: 9.3.2016.

[GeneRISK-verkkosivut, d] GeneRISK-verkkosivut. Ketkä voivat osallistua tutkimukseen. http://www.generisk.fi/node/14. Viitattu: 9.3.2016.

[GeneRISK-verkkosivut, e] GeneRISK-verkkosivut. Näytteiden, tietojen ja tutkimusten hallinta. http://www.generisk.fi/node/24. Viitattu: 9.3.2016.

[GeneRISK-verkkosivut, f] GeneRISK-verkkosivut. Tutkimuksen kulku. http://www.generisk.fi/node/28. Viitattu: 1.2.2016.

[GeneRISK-verkkosivut, g] GeneRISK-verkkosivut. Tutkimuksen tarkoitus. http://www.generisk.fi/node/20. Viitattu: 9.3.2016.

[GeneRISK-verkkosivut, h] GeneRISK-verkkosivut. Tutkittavalta kerättävät tiedot ja näytteet. http://www.generisk.fi/node/22. Viitattu: 9.3.2016.

[Honaker et al., 2011] Honaker, J., King, G., and Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of Statistical Software*, 45(7):1–47.

[Kowarik and Templ, 2016] Kowarik, A. and Templ, M. (2016). Imputation with the R package VIM. *Journal of Statistical Software*, 74(7):1–16.

[Kropko et al., 2014] Kropko, J., Goodrich, B., Gelman, A., and Hill, J. (2014). Multiple imputation for continuous and categorical data: Comparing joint multivariate normal and conditional approaches. *Political Analysis*, 22(4):497–519.

[Laaksonen, 2013] Laaksonen, S. (2013). *Surveymetodiikka - Aineiston kokoamisesta puhdistamisen kautta analyysiin, 2. painos.* bookboon.com.

[Lee and Carlin, 2010] Lee, K. J. and Carlin, J. B. (2010). Multiple imputation for missing data: Fully conditional specification versus multivariate normal imputation. *American Journal of Epidemiology*, 171(5):624–632.

[Little and Rubin, 2002] Little, R. J. and Rubin, D. B. (2002). *Statistical analysis with missind data, Second edition.* Hoboken, New Jersey: John Wiley & Sons Inc.

[Molnar et al., 2008] Molnar, F. J., Hutton, B., and Fergusson, D. (2008). Does analysis using "last observation carried forward" introduce bias in dementia research? *CMAJ*, 8(179):751–753.

[Pentala, 2014] Pentala, O. (2014). *Väestötutkimusaineiston tilastolliset kadonhallintamenetelmät*. Pro gradu -tutkielma, Helsingin yliopisto.

[Pigott, 2001] Pigott, T. D. (2001). A review of methods for missing data. *Educational Research and Evaluation*, 7(4):353–383.

[R Core Team, 2017] R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[Schafer, 1997] Schafer, J. (1997). *Analysis of Incomplete Multivariate Data (Chapman & Hall/CRC Monographs on Statistics & Applied Probability)*. Chapman and Hall/CRC.

[Taanila, 2011] Taanila, A. (2011). Ristiintaulukointi ja khiin neliö -testi. `https://tilastoapu.wordpress.com/2011/10/14/6-ristiintaulukointi-ja-khiin-nelio-testi/`. Viitattu: 18.8.2016.

[Tiirikainen, 2006] Tiirikainen, K. (2006). *Koti- ja vapaa-ajan tapaturmat: monimuuttujamenetelmät väestöryhmien välisten erojen selvittämisessä*. Pro gradu -tutkielma, Helsingin yliopisto.

[Valaste, 2015] Valaste, M. (2015). *Adjustment for Covariate Measurement Errors in Complex Surveys: A Simulation Study of Three Competing Methods*. Väitöskirja, Helsingin yliopisto.

[van Buuren, 2007] van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16:219–242.

[van Buuren and Groothuis-Oudshoorn, 2011] van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67.

[Wu et al., 2015] Wu, W., Fan, J., and Enders, C. (2015). mice: Imputation strategies for ordinal missing data, a comparison of imputation strategies for ordinal missing data on likert scale variables. *Multivariate Behavioral Research*, 50(5):484–503.

[Yu et al., 2007] Yu, L-M, B., Andrea, and Rivero-Arias, O. (2007). Evaluation of software for multiple imputation of semi-continuous data. *Statistical Methods in Medical Research 2007*, 16:243–258.

Liite A

Kysymyslomakkeen kysymysten aihepiirit

1. Kysymyssarja 1: Yleinen terveydentila
2. Kysymyssarja 2: Terveydenhuollon ammattilaisen tarve
3. Kysymyssarja 3: Verenpaineen kotimittaus
4. Kysymyssarja 4: Perheen terveyshistoria
5. Kysymyssarja 5: Liikunnan harrastaminen
6. Kysymyssarja 6: Huumeidenkäyttö
7. Kysymyssarja 7: Tupakointi
8. Kysymyssarja 8: Alkoholinkäyttö
9. Kysymyssarja 9: Nukkuminen
10. Kysymyssarja 10: Elämäntapojen hallinta
11. Kysymyssarja 11: Ruokailu

12. Kysymyssarja 12: Päivittäisistä toimista selviytyminen
13. Kysymyssarja 13: Naisten kysymykset
14. Kysymyssarja 15: Taustatiedot

Liite B

Kysymyskohtainen puuttuvuus

Taulukko B.1: Kysymyskohtainen puuttuvuus.

Kysymys	**Puuttuvat (lkm) niistä, joille kysymys suunnattu**	**Kenelle kysymys suunnattu (P=pakko vastata)**	**Puuttuvuus-prosentti**
subject	0	KAIKKI (P)	0
gender	0	KAIKKI (P)	0
dateofbirth	0	KAIKKI (P)	0
1.1A	0	KAIKKI (P)	0
1.1B	0	KAIKKI (P)	0
1.1C	0	KAIKKI (P)	0
1.1D	0	KAIKKI (P)	0
1.1E	0	KAIKKI (P)	0
1.1F	0	KAIKKI (P)	0
1.1G	0	KAIKKI (P)	0
1.1H	0	KAIKKI (P)	0
1.1I	0	KAIKKI (P)	0
1.1J	0	KAIKKI (P)	0
1.1K	0	KAIKKI (P)	0
1.1L	0	KAIKKI (P)	0
1.1M	0	KAIKKI (P)	0
1.1N	0	KAIKKI (P)	0
1.1O	0	KAIKKI (P)	0
1.1P	0	KAIKKI (P)	0
1.1Q	0	KAIKKI (P)	0
1.1R	0	KAIKKI (P)	0
1.1S	0	KAIKKI (P)	0
1.1T	0	KAIKKI (P)	0
1.1U	0	KAIKKI (P)	0
1.1V	0	KAIKKI (P)	0
1.1W	0	KAIKKI (P)	0
1.1b	0	vain niille, joilla on tällainen sairaus	0
1.1c	0	vain niille, joilla on tällainen sairaus	0
1.2	0	KAIKKI (P)	0
1.3	0	1229	0
1.4	0	678	0
1.5	0	678	0
1.6	0	678	0
1.7	18	678	2.65
1.8	5	678	0.74
1.9	2	139	1.44
1.10	7	678	1.03
1.11	4	318	1.26
1.12	0	KAIKKI (P)	0
1.13	0	1269	0
1.13b	0	640	0
1.14	1	640	0.16
1.15	0	342	0
1.16	3	342	0.88
1.17	4	640	0.63
1.18	2	261	0.77
1.19	0	KAIKKI (P)	0
1.20	1	1174	0.09
1.20b	4	1174	0.34
1.21	101	229	44.10
1.22	101	229	44.10
1.23	100	229	43.67
1.24	1	69	1.45
2.1	7	KAIKKI	0.55
2.2	20	KAIKKI	1.56
2.3	631	KAIKKI	49.37
2.4	4	KAIKKI	0.31
2.5	4	KAIKKI	0.31
2.6	0	281	0
3.1	0	KAIKKI (P)	0
3.2	0	734	0
3.3a	166	734	22.62
3.3b	167	734	22.75
3.3x	0	734	0
4.1A	0	KAIKKI (P)	0

Kysymys	Puuttuvat (lkm) niistä, joille kysymys suunnattu	Kenelle kysymys suunnattu (P=pakko vastata)	Puuttuvuusprosentti
4.1B	0	KAIKKI (P)	0
4.1C	0	KAIKKI (P)	0
4.1D	0	KAIKKI (P)	0
4.1E	0	KAIKKI (P)	0
4.1F	0	KAIKKI (P)	0
4.1G	0	KAIKKI (P)	0
4.1H	0	KAIKKI (P)	0
4.1I	0	KAIKKI (P)	0
4.1J	0	KAIKKI (P)	0
4.1K	0	KAIKKI (P)	0
4.2A	0	KAIKKI (P)	0
4.2B	0	KAIKKI (P)	0
4.2C	0	KAIKKI (P)	0
4.2D	0	KAIKKI (P)	0
4.2E	0	KAIKKI (P)	0
4.2F	0	KAIKKI (P)	0
4.2G	0	KAIKKI (P)	0
4.2H	0	KAIKKI (P)	0
4.2I	0	KAIKKI (P)	0
4.2J	0	KAIKKI (P)	0
4.2K	0	KAIKKI (P)	0
4.3	0	KAIKKI (P)	0
4.4A	0	892	0
4.4B	1	892	0.11
4.4C	0	892	0
4.4D	0	892	0
4.4E	0	892	0
4.4F	0	892	0
4.4G	0	892	0
4.4H	0	892	0
4.4I	0	892	0
4.4J	0	892	0
4.4K	2	892	0.22
4.5	0	KAIKKI (P)	0
4.6A	0	844	0
4.6B	0	844	0
4.6C	0	844	0
4.6D	0	844	0
4.6E	0	844	0
4.6F	0	844	0
4.6G	1	844	0.12
4.6H	0	844	0
4.6I	1	844	0.12
4.6J	0	844	0
4.6K	0	844	0
5.1	7	KAIKKI	0.55
5.2	5	KAIKKI	0.39
5.3	7	KAIKKI	0.55
5.4A	428	KAIKKI	33.49
5.4B	125	KAIKKI	9.78
5.4C	355	KAIKKI	27.78
5.4D	479	KAIKKI	37.48
5.4E	929	KAIKKI	72.69
6.1	3	KAIKKI	0.23
7.1	0	KAIKKI (P)	0
7.2	5	775	0.65
7.3	43	712	6.04
7.4	22	712	3.09
7.5	0	712	0
7.6	4	712	0.56
7.7	41	297	13.80
7.8A	60	297	20.20
7.8B	211	297	71.04
7.8C	224	297	75.42
7.8D	207	297	69.70
7.8b	23	297	7.74
10.15	28	297	9.43
7.9	24	297	8.08
7.10	27	297	9.09
7.11	36	297	12.12
7.12	31	297	10.44
7.13	27	297	9.09
7.14	9	297	3.03
7.15	125	KAIKKI	9.78

Kysymys	Puuttuvat (lkm) niistä, joille kysymys suunnattu	Kenelle kysymys suunnattu (P=pakko vastata)	Puuttuvuusprosentti
7.16A	242	KAIKKI	18.94
7.16B	219	KAIKKI	17.14
7.16C	224	KAIKKI	17.53
8.1	0	KAIKKI (P)	0
8.2	5	1183	0.42
8.3	11	1183	0.93
8.4	17	1183	1.44
8.5A	246	1183	20.79
8.5B	507	1183	42.86
8.5C	486	1183	41.08
8.5D	300	1183	25.36
8.5E	584	1183	49.37
8.6	17	1183	1.44
8.7	16	1183	1.35
8.8	15	1183	1.27
8.9	17	1183	1.44
8.10	20	1183	1.69
8.11	25	1183	2.11
8.12	24	1183	2.03
8.13	16	1183	1.35
9.1A	36	KAIKKI	2.82
9.1B	24	KAIKKI	1.88
9.2A	43	KAIKKI	3.36
9.2B	21	KAIKKI	1.64
9.3A	78	KAIKKI	6.10
9.3B	158	KAIKKI	12.36
9.4	6	KAIKKI	0.16
9.5	12	KAIKKI	0.94
9.6A	8	KAIKKI	0.63
9.6B	8	KAIKKI	0.63
9.6C	6	KAIKKI	0.47
9.6D	8	KAIKKI	0.63
10.1A	0	KAIKKI (P)	0
10.1B	0	KAIKKI (P)	0
10.1C	0	KAIKKI (P)	0
10.1D	0	KAIKKI (P)	0
10.1E	0	KAIKKI (P)	0

Kysymys	Puuttuvat (lkm) niistä, joille kysymys suunnattu	Kenelle kysymys suunnattu (P=pakko vastata)	Puuttuvuusprosentti
10.1F	0	KAIKKI (P)	0
10.1G	0	KAIKKI (P)	0
10.1H	0	KAIKKI (P)	0
10.1I	0	KAIKKI (P)	0
10.1J	0	KAIKKI (P)	0
10.2A	0	KAIKKI (P)	0
10.2B	0	KAIKKI (P)	0
10.2C	0	KAIKKI (P)	0
10.2D	0	KAIKKI (P)	0
10.2E	0	KAIKKI (P)	0
10.2F	0	KAIKKI (P)	0
10.5A	0	KAIKKI (P)	0
10.5B	0	KAIKKI (P)	0
10.5C	0	KAIKKI (P)	0
10.5D	0	KAIKKI (P)	0
10.5E	0	KAIKKI (P)	0
10.5G	0	KAIKKI (P)	0
10.3A	2	KAIKKI	0.16
10.3B	6	KAIKKI	0.47
10.3C	9	KAIKKI	0.70
10.3D	11	KAIKKI	0.86
10.3E	37	KAIKKI	2.90
10.3F	13	KAIKKI	1.02
10.4	0	KAIKKI	0
10.6	0	KAIKKI	0
10.7	7	KAIKKI	0.55
10.8	4	846	0.47
10.9	11	846	1.30
10.10	50	846	5.91
10.11	0	846	0
10.12	65	846	7.68
10.13	133	846	15.72
11.1A	2	KAIKKI	0.16
11.1B	9	KAIKKI	0.70
11.1C	9	KAIKKI	0.70
11.2A	25	KAIKKI	1.96
11.2B	21	KAIKKI	1.64

Kysymys	Puuttuvat (lkm) niistä, joille kysymys suunnattu	Kenelle kysymys suunnattu (P=pakko vastata)	Puuttuvuusprosentti
11.2C	9	KAIKKI	0.70
11.2D	38	KAIKKI	2.97
11.3A	1	KAIKKI	0.08
11.3B	8	KAIKKI	0.63
11.3C	15	KAIKKI	1.17
11.3D	11	KAIKKI	0.86
11.3E	4	KAIKKI	0.31
11.3F	6	KAIKKI	0.47
11.3G	6	KAIKKI	0.47
11.3H	5	KAIKKI	0.39
11.3I	7	KAIKKI	0.55
11.3J	6	KAIKKI	0.47
11.3K	4	KAIKKI	0.31
11.3L	5	KAIKKI	0.39
11.3M	7	KAIKKI	0.55
11.3N	7	KAIKKI	0.55
11.3O	5	KAIKKI	0.39
11.3P	9	KAIKKI	0.70
11.3Q	13	KAIKKI	1.02
11.3R	9	KAIKKI	0.70
11.3S	11	KAIKKI	0.86
11.3T	3	KAIKKI	0.23
11.3U	3	KAIKKI	0.23
12.1a	5	KAIKKI	0.39
12.1b	5	KAIKKI	0.39
12.1c	7	KAIKKI	0.55
12.1d	2	KAIKKI	0.16
12.1e	7	KAIKKI	0.55
12.2	7	KAIKKI	0.55
12.3	3	KAIKKI	0.23
12.4a	2	KAIKKI	0.16
12.4b	3	KAIKKI	0.23
12.4c	5	KAIKKI	0.39
12.4d	4	KAIKKI	0.31
12.4e	5	KAIKKI	0.39
12.4f	7	KAIKKI	0.55
12.4g	2	KAIKKI	0.16

Kysymys	Puuttuvat (lkm) niistä, joille kysymys suunnattu	Kenelle kysymys suunnattu (P=pakko vastata)	Puuttuvuusprosentti
12.4h	7	KAIKKI	0.55
12.4i	8	KAIKKI	0.63
12.4j	12	KAIKKI	0.94
12.4k	8	KAIKKI	0.63
12.4l	3	KAIKKI	0.23
12.5	5	KAIKKI	0.39
12.6	2	KAIKKI	0.16
12.7	3	KAIKKI	0.23
12.8	2	KAIKKI	0.16
14.2	4	KAIKKI	0.31
12.9A	5	KAIKKI	0.39
12.9B	5	KAIKKI	0.39
12.9C	4	KAIKKI	0.31
12.9D	6	KAIKKI	0.47
13.1	59	746	4.62
13.2	44	746	3.44
13.3	28	746	2.19
13.4	5	746	0.67
13.5	490	746	65.68
13.6	18	746	2.41
13.7	19	746	2.55
13.8	8	746	1.07
13.9	9	746	1.21
13.9b	110	746	14.75
13.10	11	746	1.47
13.11	5	746	0.67
15.1	36	KAIKKI	2.82
15.2	0	KAIKKI (P)	0
15.3	22	KAIKKI	1.72
15.4	36	KAIKKI	2.82
15.5	0	KAIKKI	0
15.6	0	KAIKKI	0
15.7	65	KAIKKI	5.09
15.8	36	KAIKKI	2.82
15.9	29	KAIKKI	2.27
15.10	1101	KAIKKI	86.15
15.11	0	KAIKKI	0

Kysymys	**Puuttuvat (lkm) niistä, joille kysymys suunnattu**	**Kenelle kysymys suunnattu (P=pakko vastata)**	**Puuttuvuus-prosentti**
15.12	0	KAIKKI	0

Liite C

Tutkimuksen kulku

Tutkimuksen kulku tapahtuu tutkimuskutsun lähettämisen jälkeen seuraavasti:

1. Ensin kutsun saanut rekisteröityy tutkimusportaaliin omilla pankkitunnuksilla. Rekisteröitymisen yhteydessä henkilön sopivuus tutkimukseen varmistetaan vielä kyselyllä.

2. Tutkimusportaalissa tutkittava täyttää perustietokyselylomakkeen. Tämä kyselylomake on se, jota tässä tutkielmassa käytetään.

3. Tutkittava varaa ajan terveystarkastuskäynnille, jossa tehdään perusteellinen terveystarkastus. Terveystarkastuskäynnillä hoitaja täyttää terveystietokyselylomakkeen, joka tallennetaan sähköisessä muodossa samaan tutkimusportaaliin kuin perustietokyselylomake. Vasta kun molemmat kyselylomakkeet on täytetty verkkoportaalissa, päätyvät niiden tiedot analysoitavaksi tutkimusaineistoon.

4. Tutkittava saa sydän- ja verisuonitautien riskinarvion.

5. Riskinarvion perusteella tutkittava voidaan ohjata lääkärin hoitoon. Lisäksi jokainen tutkittava voi osallistua sähköiseen elintapavalmennukseen.

6. Noin kahden vuoden kuluttua ensimmäisestä kutsusta tutkittava kutsutaan seurantatutkimukseen. [GeneRISK-verkkosivut, f]

Liite D

Kutsukirje

KUTSUKIRJE

[Kaupunki] [Päivämäärä]

[Etunimi] [Sukunimi]
[Osoite]
[Postinumero] [Postitoimipaikka]

KUTSU TERVEYSTUTKIMUKSEEN

Hyvä [Etunimi Sukunimi],

Kutsumme teidät osallistumaan Helsingin yliopiston ja Kymenlaakson sairaanhoito- ja sosiaalipalvelujen kuntayhtymän Carean toteuttamaan GeneRISK-terveystutkimukseen. Tämän tutkimuksen tavoitteena on tutkia, voidaanko sydän- ja verisuonitauteja ehkäistä hyödyntämällä sellaista tietoa perimästä, jolla on merkitystä näiden sairauksien synnyssä. Tutkittavalle tehdään tutkimuksen puitteissa maksuton terveystarkastus ja palautetaan tutkimuksessa annettujen tietojen perusteella laskettu sydän- ja verisuonitautien riskiarvio. Tietoa tutkimuksen yksityiskohdista saatte oheisesta Tutkittavan tiedotteesta ja Tiedotteesta biopankkinäytteen antajalle. Tutkimus koostuu kahdesta osasta: 1) tutkimuspaikalla tehtävästä terveystarkastuksesta, jonka yhteydessä otetaan verinäyte, ja 2) täytettävästä kyselylomakkeesta.

1) Terveystarkastus

Mikäli haluatte osallistua tutkimukseen, teidät kutsutaan maksuttomalle terveystarkastuskäynnille osoitteeseen:

Kymenlaakson keskussairaala, päärakennus, 6. krs, huoneet 1 ja 2
Kotkantie 41
48210 Kotka

Pyydämme teitä varaamaan ajan terveystarkastukseen mahdollisimman pian soittamalla numeroon 020 633 2733 arkipäivänä klo 12.45-14.45[1]. Tutkimukseen otetaan ilmoittautumisjärjestyksessä. Mikäli teillä on kysyttävää tutkimuksesta tai teille annettavasta tiedosta, soittakaa numeroon 020 633 2733 arkipäivisin klo 12.45-14.45 välillä.

Tutkimus on teille ilmainen. Valitettavasti emme pysty korvaamaan tutkimuspaikalle saapumisesta aiheutuvia kuluja.

<u>Valmistautuminen terveystarkastukseen</u>

Pyydämme teitä saapumaan tutkimuspaikalle viimeistään 5-10 minuuttia ennen yllä mainittua tutkimusaikaanne ja varaamaan tutkimukseen aikaa noin 45 minuuttia. Mikäli ette ole täyttäneet alla mainittua kyselylomaketta ennen tutkimukseen saapumista, pyydämme teitä saapumaan viimeistään 45 minuuttia ennen tutkimusaikaanne, jotta ehditte täyttää lomakkeen ennen terveystarkastusta. Terveystarkastusta ja näytteenottoa ei voida tehdä ennen kuin kyselylomake on täytetty.

[1] Palvelu on maksuton. Soittaja maksaa puhelusta teleoperaattorille. Puhelun hinta määräytyy seuraavilla periaatteilla: 1) suomalaisesta lankaliittymästä soitettaessa peritään paikallisverkkomaksu (pvm), 2) suomalaisen puhelinyhtiön matkapuhelinliittymästä Suomessa soitettaessa peritään matkapuhelinmaksu (mpm), 3) ulkomailta soitettaessa hinnan määrittelee paikallinen operaattori. Puhelujen tarkka hinta määräytyy kuitenkin asiakkaan puhelinyhtiön kanssa tekemän liittymäsopimuksen perusteella. Operaattoreiden erilaisissa puhepaketeissa on yleensä rajattu pois soitot yrityspuhelinnumeroihin, jollaisia ovat mm. 020 6-alkuiset puhelinnumerot. Näistä pakettiin kuulumattomista soitoista puhelinyhtiö perii puhepakettisopimuksensa mukaisen normaalin puhelumaksun (mpm/pvm). Puhelinyhtiöt kertovat hinnoitteluperusteensa asiakkailleen tarkemmin puhepaketin liittymäsopimuksessa.

Luotettavien veren rasva- ja sokeriarvojen saamiseksi teidän tulisi olla syömättä ja juomatta 10 tuntia ennen terveystarkastusta (edes purukumin syönti ei ole sallittua). Vettä voi juoda pieniä määriä. Ravinnotta olo ei koske tyypin 1 (nuoruustyypin) diabeetikkoja.

Ottakaa ystävällisesti terveystarkastukseen mukaan henkilöllisyystodistus, jotta voimme varmistaa henkilöllisyytenne. Ottakaa mukaan myös kaikki säännöllisesti käyttämienne lääkkeiden purkit, jotta voimme kirjata näistä tutkimuksen kannalta tärkeät tiedot. Toivomme teidän pukeutuvan niin, että voitte vaivatta riisua oikean olkavartenne paljaaksi verenpaineen mittausta varten. Pituuden ja painon mittausta varten riisutaan kengät, ja vyötärön- ja lantionympäryksen mittausta varten paksut tai kiristävät vaatteet.

Pyydämme teitä myös tutustumaan etukäteen liitteenä olevaan Tutkittavan tiedotteeseen ja Tiedotteeseen biopankkinäytteen antajalle sekä suostumuslomakkeeseen. Ottakaa ystävällisesti suostumuslomake mukaanne tutkimuspaikalle. Tutkimuspaikalla teiltä pyydetään kirjallinen suostumus tutkimustietojenne käyttöön.

2) Kyselylomake

Ennen terveystarkastukseen saapumista pyydämme teitä täyttämään henkilökohtaisen kyselylomakkeenne internetissä osoitteessa my.generisk.fi. Lomakkeen täyttö kestää keskimäärin 30 minuuttia. Lomakkeen täytön voi tarvittaessa keskeyttää ja täyttöä voi jatkaa myöhemmin henkilökohtaisilla tunnuksilla.

Pyydämme teitä täyttämään kyselyn mahdollisimman huolellisesti lomakkeessa annettujen ohjeiden mukaisesti. Mikäli lomakkeessa on kohtia, joihin vastaaminen on vaikeaa, voitte jättää ne avoimiksi ja keskustella niistä tutkimushoitajien kanssa tutkimuspaikalla.

Mikäli teillä ei ole mahdollisuutta täyttää kyselyä internetissä, pyydämme ilmoittamaan asiasta soittamalla numeroon 020 633 2733 (arkipäivisin klo 12.45-14.45), jotta voimme lähettää teille paperilomakkeen. Paperilomakkeella kyselyyn vastanneita pyydetään ottamaan pankkitunnuksensa mukaan terveystarkastukseen, jotta he voivat rekisteröityä tutkimukseen tutkimushoitajan avustuksella (pankkitunnuksia käyttää ainoastaan tutkittava itse, tutkimushoitaja ei käsittele eikä näe pankkitunnuksia).

Tutkimuksessa noudatetaan tietojen käsittelyn suhteen Henkilötietolakia, Lakia potilaan asemasta ja oikeuksista ja muita asiaan liittyviä lakeja ja säädöksiä.

Kunnioittaen,

Ylilääkäri Pasi Pöllänen, LT, FT, dosentti
Carea – Kymenlaakson sairaanhoito- ja sosiaalipalvelujen kuntayhtymä

Vanhempi tutkija Elisabeth Widén, LKT, dosentti
Suomen molekyylilääketieteen instituutti FIMM, Helsingin yliopisto