Tilastotieteen pro gradu -tutkielma

Riskitiheyskuvio Coxin mallin diagnostiikassa ja tulkinnassa sekä sen toteuttaminen R-ympäristöön

Nanni Koski

Jyväskylän yliopisto
Matematiikan ja tilastotieteen laitos
29. tammikuuta 2015

Kiitokset

Haluan kiittää useita henkilöitä tutkielmani valmistumisen edistämisestä. Ensinnäkin kiitän ohjaajaani professori Juha Karvasta kiinnostavasta aiheesta sekä ohjauskeskusteluista. Osoitan kiitokseni myös Gerontologian tutkimuskeskukselle Ikivihreät-aineiston antamisesta käyttööni sekä erityisesti Markku Kauppiselle ja Tiina-Mari Lyyralle.

Kiitän Salme Kärkkäistä ja Jouni Helskettä lähdekirjallisuuden lainasta. Lisäksi Helske ansaitsee kiitokset avusta R-paketin kokoamisessa ja CRANiin saattamisessa.

Jaakko Reinikainen oli avuksi tekemäni koodin testaamisessa. Kiitän häntä myös funktioiden kehitysideoista sekä LaTeX-avusta. Louis Planelle kuuluu kiitos englannin kielen oikoluvusta sekä kieliopillisten seikkojen tarkistamisesta. Näkemyksistä sekä keskusteluista kiitän Johanna Ärjeä, Anna-Kaisa Ylitaloa ja Satu Helskettä.

Kiitokset myös ystävilleni sekä perheelleni kaikesta avusta ja tuesta.

Tiivistelmä

Tutkielmassa esitellään riskitiheyskuvio, jota voidaan käyttää Coxin suhteellisten riskitiheyksien mallin tulosten sekä aineiston havainnollistamisessa. Kaikki mallissa olevat kovariaatit voidaan piirtää samaan kuvaan, sillä vaaka-akselin koordinaattina käytetään $[0, 1]$-välille skaalattuja järjestyslukuja. Pystyakselille piirretään kovariaateittain suhteellinen riskitiheys, joka on mallin antama riskitiheys jaettuna riskitiheydellä vertailukohdassa kuten kovariaatin mediaanissa.

Riskitiheyskuvio mahdollistaa mallissa olevien kovariaattien keskinäisen merkityksen arvioinnin populaatiotasolla. Riskitiheyskuviota voidaan käyttää myös kovariaatin muunnosten tulkinnalliseen vertailuun. Kuvio näyttää mallin antamat suhteelliset riskitiheydet kovariaattien minimissä ja maksimissa, ja sen avulla voidaan siis arvioida, tuottaako suhteellisten riskitiheyksien oletus järkeviä riskitiheyden estimaatteja havaintoyksiköille, joilla on havaittu äärimmäisiä arvoja kovariaateilla. Lisäksi esitetään esimerkki riskitiheyskuvion hyödyntämisestä Coxin mallille, jossa on ajassa muuttuvia kovariaatteja.

Riskitiheyskuvio voidaan piirtää R-ympäristön rankhazard-pakettia käyttäen. Tutkielman pääpaino on ollut rankhazard-paketin päivittämisessä aiemmasta versiosta 0.8-1. Tutkielman tuloksena paketista julkaistiin R-ohjelmapakettien kokoelmaan (The Comprehensive R Archive Network) versio 1.0. Siihen on korjattu edellisen version puutteita sekä lisätty uusia ominaisuuksia. Nyt malleissa voi käyttää sekä faktoreita että muunnoksia. Lisäksi suhteellisen riskitiheyden luottamusvälien piirtäminen on mahdollista, ja muun muassa graafisia ominaisuuksia on parannettu lisäargumenttien myötä.

Tekijä: Koski, Nanni
Työn nimi: Riskitiheyskuvio Coxin mallin diagnostiikassa ja tulkinnassa sekä sen toteuttaminen R-ympäristöön
Tieteenala: Tilastotiede
Sivumäärä: 48 s. + liitteet 75 s.
Korkeakoulu: Jyväskylän yliopisto
Laitos: Matematiikan ja tilastotieteen laitos
Valmistumisaika: Tammikuu 2015

Sisältö

1 Johdanto

Edward Tufte (1983) on esittänyt suuntaviivoja laadukkaalle aineiston graafiselle esittämiselle. Graafisen esityksen tulee olla selkeä ja tarkoituksenmukainen. Laadukas kuvaaja esittää aineiston totuudenmukaisesti, ja se saa ajattelemaan esitetyn asian sisältöä, ei kuvaajan tuottamista tai suunnittelua. Erinomainen kuvaaja sisältää myös paljon aineistoa pienessä tilassa, ja se antaa katselijalle nopeasti informaatiota.

Tässä työssä käsitellään graafista esittämistä elinaika-analyysissä. Elinaikamallissa vasteena on aika tietyn seurattavana olevan tapahtuman ilmenemiseen (Lee & Wang, 2003). Tämä tapahtuma voi olla esimerkiksi laitteen rikkoontuminen, kuolema, sairastuminen, maanjäristys, auto-onnettomuus, pidätetyksi tuleminen, lapsen syntymä, työn saaminen, ylennyksen saaminen, irtisanominen, eläkkeelle jääminen tai avioero (Allison, 1995). Aika voidaan mitata sekunneissa, minuuteissa, päivissä, vuosissa; yksiköllä ei ole merkitystä (Cleves, Gould, Gutierrez & Marchenko, 2008). Lähtökohta elinaika-analyysiin on stokastinen, mikä tarkoittaa, että henkilön havaittu elinaika t ajatellaan havainnoksi ei-negatiivisen satunnaismuuttujan T jakaumasta (Allison, 1995).

Tilastollisiin malleihin liittyvät graafiset esitykset perustuvat usein mallin jäännöksiin tai jäännösten muunnoksiin, kuten standardoituihin residuaaleihin. Jäännöskuvioiden avulla nähdään, sopiiko malli aineistoon, mutta ne antavat harvoin informaatiota alkuperäisestä aineistosta tai sovitetun mallin tuloksista. (Therneau & Grambsch, 2000.)

Yksi käytetyimmistä malleista elinaika-analyysissä on Coxin suhteellisten riskitiheyksien malli (Allison, 1995). Karvanen ja Harrell ovat esitelleet vuonna 2009 artikkelissaan *Visualizing covariates in proportional hazards model* riskitiheyskuvion, jolla visualisoidaan Coxin mallin tuloksia sekä käytettyä aineistoa. Riskitiheyskuviolla (engl. *rank-hazard plot*) havainnollistetaan mallissa olevien kovariaattien yhteyttä suhteellisen riskitiheyden suuruuteen, ja kuvion perusteella voidaan tehdä mallista sisällöllisiä tulkintoja.

Riskitiheyskuvioon piirretään Coxin mallilla estimoidut suhteellisen riskitiheyden jakaumat kovariaateittain, mikä mahdollistaa kovariaattien välisen vertailun. Suhteellinen riskitiheys on mallin antama riskitiheys jaettuna valitun vertailukohdan riskitiheydellä. Riskitiheyskuvio täydentää muita Coxin mallin diagnostiikkakeinoja, ja sen avulla voidaan vastata kysymyksiin, joita muut diagnostiikkakeinot eivät käsittele. Erityisesti voidaan arvioida populaatioriskiä eri kovariaateilla, ja vertailla kovariaatteja populaatiotasolla. (Karvanen & Harrell, 2009.)

Riskitiheyskuvion avulla voidaan tarkastella mallin antamien tulosten järkevyyttä. Kuviosta nähdään suhteellisen riskitiheyden arvot kunkin kovariaatin minimissä ja maksimissa. Karvasen ja Harrellin (2009) mukaan kuvion avulla voidaan päätellä, tuottaako Coxin mallin oletus suhteellisista riskitiheyksistä järkeviä estimaatteja yksilöille, joilla selittävän muuttujan arvot ovat joko erittäin suuria tai pieniä. On nimittäin mahdollista, että malli vaikuttaa sopi-

van hyvin aineistoon, mutta mallin antamat estimaatit joillekin kovariaattien arvoille ovat epärealistisia.

Jotta riskitiheyskuvion käyttö voi yleistyä, tulee sen olla helposti saatavilla johonkin tilastolliseen ohjelmistoon. R-ympäristö (R Core Team, 2014) mahdollistaa käyttäjiensä tekemät laajennusosat eli paketit. Mikäli paketti täyttää tietyt vaatimukset, se voidaan ladata osaksi R-ohjelmapakettien kokoelmaa (The Comprehensive R Archive Network, lyh. CRAN), jolloin kaikki R-ympäristön käyttäjät voivat ladata paketin käyttöönsä. Vaatimuksiin kuuluvat mm. koodin toimivuus sekä paketin kautta käytettävien funktioiden dokumentointi.

Juha Karvanen on luonut `rankhazard`-paketin R-ympäristön laajennusosaksi vuonna 2009. Paketissa olevalla `rankhazardplot`-funktiolla piirretään riskitiheyskuvioita. Paketin ensimmäinen versio oli numeroltaan 0.8. Karvanen on myös päivittänyt pakettia versioksi 0.8-1 vuonna 2012 (Karvanen, 2012). Pro gradu -työni tavoitteena on ollut `rankhazard`-paketin kehittäminen sekä riskitiheyskuvion käyttämisen esittely. Tekemäni versio 1.0 `rankhazard`-paketista on julkaistu CRANiin 2.4.2014 (Karvanen & Koski, 2014). Tällöin liityin `rankhazard`-paketin toiseksi tekijäksi.

Keskeisintä paketin kehittämisessä on ollut luoda toimiva ja helppokäyttöinen versio, jotta riskitiheyskuvio voi hyödyttää tiedeyhteisöä. Uusimmassa versiossa on mahdollista piirtää riskitiheyskuvioita malleista, joissa selittävinä muuttujina on faktoreita ja muunnoksia. Muita pääsaavutuksia ovat luottamusvälien piirtämisen mahdollisuus, graafisten ominaisuuksien monipuolistaminen ja korjaaminen sekä piirrettävien suhteellisen riskitiheyden jakaumien valinta mallista. Olen myös ideoinut tavan käyttää riskitiheyskuviota Coxin regressiossa ajassa muuttuvilla kovariaateilla.

Tutkielmassa esitellään riskitiheyskuvion hyödyntämistä diagnostiikassa sekä tulkinnassa ja tarjotaan käyttöohjeet riskitiheyskuvioiden piirtämiseen `rankhazardplot`-funktiolla. Tutkielma tähtää riskitiheyskuvion ominaisuuksien esittelyyn, ei täysimittaiseen data-analyysiin. Tästä syystä mallien sopivuutta aineistoon ei tarkastella. Esimerkeissä käytetyt mallit on sovitettu, jotta riskitiheyskuvion piirtämistä sekä tulkintaa voidaan havainnollistaa mahdollisimman monipuolisesti.

Luvussa 2 luodaan katsaus elinaika-analyysiin yleisesti sekä Coxin malliin. Lisäksi käydään läpi Coxin mallin diagnostiikkakeinoja. Luvussa 3 esitellään R-ympäristöä sekä käytettäviä aineistoja. R-ympäristön peruskäytön oletetaan olevan lukijan hallinnassa. Riskitiheyskuvion piirtämiseen ja tulkintaan perehdytään luvussa 4. Luvussa esitellään myös useita esimerkkejä riskitiheyskuvion käytöstä. Luvussa 5 katsotaan, kuinka riskitiheyskuviota voi hyödyntää Coxin mallin diagnostiikassa, kun taas luvussa 6 käydään läpi `rankhazard`-paketin käyttämistä sekä riskitiheyskuvion piirtämistä R-koodin kautta. Lukuun 6.3 on kirjattu pakettiin tekemäni muutokset. Lopuksi ennen liitteitä on pohdinta.

Liitteet tarjoavat teknistä lisätietoa riskitiheyskuvion piirtämisestä. Liite A

sisältää yksityiskohtaisen esittelyn `rankhazardplot`-funktion toiminnasta sekä argumenttien käytöstä. Liite B on Internetiin julkaistu dokumentaatio `rankhazardplot`-funktiolle. Dokumentaation esimerkit on tulostettu liitteeseen C. Liite D sisältää `rankhazard`-paketin version 1.0 koodit. Liitteessä E on koodi, jota käyttäen on luotu aineisto, jolla voidaan sovittaa Coxin malli ajassa muuttuvilla kovariaateilla. Sitä aineistoa käytetään luvussa 4.5 sovitetussa mallissa.

2 Elinaika-analyysistä

Tässä luvussa tutustutaan elinaika-analyysin erityispiirteisiin ja esitellään elinaikamallinnuksen kannalta kiinnostavia funktioita. Käydään myös läpi tutkielmassa käytettävää Coxin mallia, ja luodaan katsaus sen diagnostiikkakeinoihin.

2.1 Elinaikavasteen sensurointi

Elinaikamallin vaste on määrätystä alkuhetkestä johonkin tapahtumaan kuluva aika. Seurattavana oleva tapahtuma ei aina ilmene tutkimusaikana tai joissain tapauksissa ei tiedetä tarkkaa aikaa, jolloin tapahtuma on ilmennyt. Tätä kutsutaan sensuroinniksi. Sensurointi voi tapahtua vasemmalta, oikealta tai jollakin välillä. Jos henkilön i elinaika on vasemmalta sensuroitu, tiedetään, että $T_i < c$, missä c on tunnettu. Vastaavasti, jos elinaika T_i on oikealta sensuroitu, tiedetään, että $T_i > c$. Jos taas elinaika on sensuroitu jollain välillä, tiedetään $c_\mathrm{a} < T_i < c_\mathrm{y}$, missä c_a ja c_y ovat tunnetut ala- ja ylärajat elinajalle. (Allison, 1995.)

Esimerkiksi tentin tekemiseen käytetty aika on vasemmalta sensuroitu henkilöille, jotka lähtevät tentistä heti vähimmäisajan jälkeen. Elinaika tutkimuksessa on oikealta sensuroitu henkilöille, jotka ovat elossa vielä tutkimuksen päättyessä. Aika, joka kuluu siihen, että kala jää katiskaan, on välisensuroitu. Alaraja on tällöin katiskan edellinen tarkastamisaika ja yläraja on kalan sisältävän katiskan nostamisaika. Tarkkaa hetkeä, jolla kala ui katiskaan, ei tiedetä.

Ajan kesto mitataan alkuhetkestä tapahtumaan tai yksilön sensuroitumiseen asti. Alkuhetki voi olla esimerkiksi henkilön syntymä, diagnoosin saaminen, ammattiin valmistuminen tai avioliiton solmiminen. Vaikka vaste ei siis välttämättä ole suoranaisesti elinaika, kutsutaan sitä tästä eteenpäin elinajaksi, ja seurattavaa tapahtumaa kuolemaksi, ellei muuten määritellä. Elinaika-analyysissä vasteen määrittelyssä tulee tietää sekä elinaika että sen status, joka kertoo, onko elinaika havaittu vai sensuroitu.

2.2 Välttö- ja riskitiheysfunktio

Jatkuvan elinajan T tuntematonta todennäköisyysjakaumaa voidaan kuvata sekä tiheysfunktion f että kertymäfunktion F avulla. Kertymäfunktio F kertoo todennäköisyyden elää vähintään aika t, siis $F(t) = P(T \leq t)$. Kun vasteena on elinaika, on kuitenkin usein sopivampaa käsitellä välttöfunktiota S, joka määritellään todennäköisyytenä, että elinaika on suurempi kuin t (Cleves ym., 2008). Välttöfunktion yhteys tiheys- ja kertymäfunktioon on:

$$S(t) = P(T > t) = \int_t^\infty f(u)\,du = 1 - \int_0^t f(u)\,du = 1 - F(t).$$

Välttöfunktio on vähenevä, ja sille pätee $S(0) = 1$ sekä $S(t) \to 0$, kun $t \to \infty$.

Myös tiheysfunktiolle on olemassa elinaika-analyysin kannalta kiinnostavampi vastine: riskitiheysfunktio määrittää todennäköisyyden kuolla jollain hetkellä sillä ehdolla, että on elänyt sen alkuun asti. Riskitiheysfunktion matemaattinen muotoilu ja yhteys kertymä- ja tiheysfunktioon on:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T < t + \Delta t \,|\, T > t)}{\Delta t} = \frac{f(t)}{1 - F(t)}.$$

Siis riskitiheysfunktion ero tiheysfunktioon on ehdollistaminen hetkeen t. Tiheys-, kertymä-, riskitiheys- sekä välttöfunktio määrittelevät saman todennäköisyysjakauman satunnaismuuttujalle T. (Cleves ym., 2008.)

Riskitiheys ei voi olla pienempi kuin nolla, mutta sillä ei ole ylärajaa. Riskitiheys kannattaa ajatella yksilön ominaisuutena, jolloin jokaisella havaintoyksiköllä voi olla oma riskinsä kuolla tietyllä hetkellä. Mikäli yksilöt ovat samanlaisia, niillä voi olla sama riskitiheysfunktio. (Allison, 1995.)

Elinaika-analyysissä kiinnostus kohdistuu nimenomaan siihen, mikä on kuolemisen riski minä tahansa ajankohtana tutkimuksen alun jälkeen. Tästä syystä mallinnetaan riskitiheyttä. Riskitiheyttä voidaan tutkia kahdesta eri näkökulmasta, jotka ovat riskitiheyttä selittävät muuttujat tai riskitiheyden estimaatti. (Collett, 2003.)

Seuraavaksi tutustutaan Coxin malliin, jonka avulla voidaan selvittää riskitiheyttä selittäviä muuttujia ilman, että riskitiheysfunktiota tarvitsee estimoida.

2.3 Coxin malli ja osittaisuskottavuus

Coxin suhteellisten riskitiheyksien mallilla on useita kutsumanimiä kuten Coxin malli, suhteellisten riskitiheyksien malli, suhteellisten intensiteettien malli, verrannollisten riskitiheyksien malli, verrannollisten intensiteettien malli (Läärä, Luostarinen, Hakulinen, Lyytikäinen, Sarna, Virtala, Riihimäki & Hakama, `http://www.finepi.org/files/englantisuomi.pdf`, 2008). Englanniksi mallia kutsutaan nimellä the (Cox) proportional hazards model, Cox regression tai lyhyesti Cox model. Tässä työssä käytetään nimiä Coxin malli ja (Coxin) suhteellisten riskitiheyksien malli.

Coxin suhteellisten riskitiheyksien mallilla voidaan tutkia eri selittäjien eli kovariaattien yhteyttä riskitiheyteen. Riskitiheysfunktio saadaan estimoitua jokaiselle havaintoyksikölle. Mallin nimi tulee siitä, että siinä oletetaan riskitiheyksien suhteen kahden havaintoyksikön välillä pysyvän samana koko tarkasteluajan. Siis kahden havaintoyksikön i ja j välillä laskettu riskitiheyksien suhde

$$\frac{h_i(t)}{h_j(t)}.$$

on vakio, ekä siisi muutu ajan t funktiona. (Collett, 2003.)

Coxin malli on nimetty Sir David Roxbee Coxin mukaan. Hän esitteli mallin artikkelissaan *Regression Models and Life-Tables* vuonna 1972 (Cox, 1972).

Mallin lisäksi Cox esitteli uuden estimointimenetelmän, jota myöhemmin on ryhdytty kutsumaan osittaisuskottavuudeksi (*partial likelihood*) tai tarkemmin suurimmaksi osittaisuskottavuudeksi (*maximum partial likelihood*) (Allison, 1995).

Esitetään seuraavaksi Coxin malli, kun tutkimuksen alussa on mitattu p selittävää muuttujaa eli kovariaattia. Luvussa 4.5 mallille on esitetty yleistys, jossa selittävien muuttujien arvot saavat muuttua seurannan aikana.

Olkoon $x_{i,j}$ henkilön i kovariaatin j arvo, ja $\boldsymbol{x}_i = (x_{i,1} \dots x_{i,p})'$ henkilön i kaikkien kovariaattien arvot vektorina. Merkitään perusriskitiheysfunktiota $h_0(t)$. Perusriskitiheys on riskitiheys henkilölle, jolla kaikkien kovariaattien arvo on nolla. Tällöin jokaiselle henkilölle $i = 1, \dots, n$ riskitiheysfunktio määritellään (Collett, 2003):

$$h_i(t) = h_0(t) \exp(\beta_1 x_{i1} + \dots + \beta_p x_{ip}) = h_0(t) \exp(\boldsymbol{\beta}' \boldsymbol{x}_i), \tag{1}$$

missä $\boldsymbol{\beta} = (\beta_1 \dots \beta_p)'$ on selittävien muuttujien kerroinvektori mallissa. Perusriskitiheysfunktio kuvaa, kuinka riski muuttuu ajassa, ja $\exp(\boldsymbol{\beta}' \boldsymbol{x}_i)$ kuvaa kovariaattien yhteyttä riskitiheyteen (Lee & Wang, 2003). Nämä mallin komponentit voidaan estimoida erikseen, eikä siksi perusriskitiheyttä tarvitse määritellä, jotta voidaan tutkia riskitiheyttä selittäviä muuttujia (Collett, 2003).

Kertoimet $\boldsymbol{\beta}$ estimoidaan numeerisesti osittaisuskottavuusfunktion avulla. Osittaisuskottavuus on

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{\exp(\boldsymbol{\beta}' \boldsymbol{x}_{(j)})}{\sum_{l \in R(t_{(j)})} \exp(\boldsymbol{\beta}' \boldsymbol{x}_l)}, \tag{2}$$

missä r on seuranta-aikana kuolleiden määrä, $\boldsymbol{x}_{(j)}$ on j:ntenä kuolleen henkilön kovariaattien arvot, $t_{(j)}$ on j:ntenä kuolleen henkilön elinaika ja $R(t_{(j)})$ on riskijoukko hetkellä $t_{(j)}$ eli ne henkilöt, jotka ovat elossa ja sensuroimattomia hetkellä $t_{(j)}$. (Collett, 2003.)

Sama voidaan esittää koko aineistolle ottaen oikealta sensurointi eri tavalla huomioon. Muuttuja δ_i saa arvon 1, jos henkilö i on kuollut seuranta-aikana ja arvon 0, jos henkilö on sensuroitunut. Nyt kaava (2) voidaan esittää myös muodossa

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left\{ \frac{\exp(\boldsymbol{\beta}' \boldsymbol{x}_i)}{\sum_{l \in R(t_i)} \exp(\boldsymbol{\beta}' \boldsymbol{x}_l)} \right\}^{\delta_i}.$$

Osittaisuskottavuuden kaavasta (2) nähdään, että elinajat tulee voida järjestää suuruusjärjestykseen. Elinaika ajatellaan jatkuvaksi muuttujaksi, jolloin kaikki elinajat ovat erisuuria. Mittaustarkkuuden vuoksi aineistoissa esiintyvät elinajat voivat kuitenkin olla yhtä suuria. Vaikka elinajat esimerkiksi päivissä mitattuna olisivat samat useilla henkilöillä, he eivät ole kuitenkaan kuolleet juuri yhtä aikaa. Elinajoilla on olemassa järjestys, joka ei ole tiedossa. Tästä syystä kaavan (2) osittaisuskottavuuden käyttäminen suoraan ei ole mahdollista, sillä kaikkia elinaikoja ei voida laittaa suuruusjärjestykseen. (Lee & Wang, 2003.)

Osittaisuskottavuuden käyttäminen on mahdollista, kun otetaan huomioon kaikki mahdolliset järjestykset, joissa samana hetkenä havaitut elinajat ovat voineet tapahtua. Jos esimerkiksi tietyllä ajanhetkellä on kuollut 5 henkilöä, mahdollisia kuolemisjärjestyksiä on yhteensä $5! = 120$. Kaikilla eri järjestysmahdollisuuksilla on oma todennäköisyys, ja niiden yhdiste tulee sisällyttää osittaisuskottavuuteen. Mikäli aineistossa on paljon sidoksia, tulee uskottavuusfunktion laskemisesta työlästä. Siksi uskottavuusfunktiolle on kehitetty approksimaatioita. Yksi näistä approksimaatioista on Efronin (1977) approksimaatio, jota on käytetty tämän tutkielman mallien sovituksessa. Approksimaatiot toimivat hyvin, kun sidosten osuus riskijoukossa kullakin havaitulla elinajalla on pieni. (Allison, 1995.)

Sen lisäksi että tietyllä hetkellä useiden henkilöiden kuoleminen on mahdollista, voi samalla hetkellä esiintyä sensuroituja havaintoja. Tällöin ajatellaan sensuroinnin tapahtuvan kuolemien jälkeen. Sensuroidut havainnot eivät siis vaikeuta riskijoukon määrittämistä kullakin ajanhetkellä. Riittää, että uskottavuusfunktiossa kyetään käsittelemään sidokset havaituissa elinajoissa. (Collett, 2003.)

2.3.1 Faktorin käyttäminen mallissa

Tässä työssä kutsutaan faktoriksi laadullista muuttujaa, jonka arvojoukko on lueteltavissa. Faktorin arvoja kutsutaan tasoiksi. Esimerkiksi sukupuoli on faktori, ja sen tasot ovat mies ja nainen. Faktorilla voidaan myös määritellä esimerkiksi siviilisääty tai koulutustaso. Kaikkien faktoreiden tasot voidaan koodata numeerisiksi.

Tarkastellaan tilannetta, jossa mallinnetaan yhden muuttujan, faktorin A, vaikutusta riskitiheyteen. Olkoon faktorin A tasojen lukumäärä a, ja olkoon β_{A_k} faktorin A vaikutusta kuvaava termi yksilölle i, jolle faktorin A taso on k. Tällöin Coxin mallin mukainen riskitiheys on $h_i(t) = h_0(t)\exp(\beta_{A_k})$. (Collett, 2003.)

Perusriskitiheysfunktio $h_0(t)$ on määritelty riskitiheydeksi henkilölle, jonka kaikkien selittävien muuttujien arvot ovat nollia. Yhdenmukaisuuden vuoksi faktorille määritellään vertailutaso, jolle kerroin on nolla. Esimerkiksi ensimmäisen tason kerroin faktorille A, siis $\beta_{A_1} = 0$. Tällöin perusriskitiheysfunktio on riskitiheys yksilölle, jolle faktorin A taso on 1. (Collett, 2003.)

Faktori A voidaan sisällyttää Coxin mallin lineaariseen osaan esittämällä se indikaattorimuuttujien X_{A_k} avulla. Indikaattorimuuttuja X_{A_k} saa arvon 1, jos faktorin taso on k, muuten arvon 0. Indikaattorimuuttujia tarvitaan $a - 1$ kappaletta, yksi jokaiselle vertailutasosta eroavalle tasolle. Mikäli faktorin taso on vertailutaso, jokaisen indikaattorimuuttujan arvo on 0. Tällöin β_{A_k} voidaan esittää muodossa $\beta_{A_2}x_{i,A_2} + \beta_{A_3}x_{i,A_3} + \ldots + \beta_{A_a}x_{i,A_a}$, missä x_{i,A_l} on indikaattorimuuttujan X_{A_l} arvo yksilölle i, $l = 2, \ldots, a$.

Yksinkertaisin esimerkki on kaksitasoinen faktori, esimerkiksi sukupuoli. Alkuperäinen muuttuja saa arvoja ”mies” ja ”nainen”. Olkoon ”mies” vertailutaso. Tällöin syntyy yksi uusi indikaattorimuuttuja, joka saa arvon 1, kun

havaintoyksikkö on nainen ja arvon 0, kun havaintoyksikkö on mies.

2.3.2 Luottamusvälien estimointi Coxin mallin kertoimille

Kun Coxin mallin kertoimet estimoidaan numeerisesti, saadaan samalla niiden keskivirheet. Estimoidun kertoimen $\hat{\beta}$ ja sen keskivirheen $\mathrm{se}(\hat{\beta})$ avulla voidaan laskea β-kertoimelle $100(1-\alpha)$ prosentin luottamusväli. Useimmiten esitetään 95 prosentin luottamusväli, jolloin $\alpha = 0.05$.

Luottamusväli saadaan kaavalla $\hat{\beta} \pm z_{\alpha/2}\,\mathrm{se}(\hat{\beta})$, missä $z_{\alpha/2}$ on standardinormaalijakauman $(1-\alpha/2)$-fraktiili. Fraktiili on arvo, johon mennessä on kertynyt kyseessä oleva osuus jakauman pinta-alasta. Standardinormaalijakauman 0.975-fraktiili on 1.96, ja sen avulla saadaan 95 prosentin luottamusväli. Mikäli lasketulle luottamusvälille ei sisälly lukua 0, se antaa näyttöä siitä, että kerroin β eroaa nollasta. (Collett, 2003).

2.4 Coxin mallin diagnostiikkamenetelmiä

Sovitetun mallin riittävyyttä tulee tarkastella eri näkökulmista. Mallin tulee esimerkiksi sisältää sopiva osajoukko mitatuista muuttujista. Coxin malli pohjautuu oletukseen suhteellisista riskitiheyksistä, ja on tärkeää tutkia myös tämän oletuksen voimassaoloa. (Collett, 2003.)

Seuraavaksi käydään läpi joitakin diagnostiikkamenetelmiä Coxin mallille. Esiteltävät menetelmät on pääasiassa valittu niiden yleisyyden perusteella: menetelmät mainitaan useissa lähdeteoksissa, ja ne ovat saatavilla eri ohjelmistoissa. Tässä luvussa on tarkoitus antaa käsitys siitä, millaisia ja mihin tarkoituksiin sopivia menetelmiä on olemassa. Muunnosten ja vaikuttavien havaintojen diagnostiikkaa käsitellään luvussa 5.

Coxin mallille on kehitetty useita erilaisia residuaaleja. Residuaalien avulla pyritään tutkimaan, ovatko poikkeamat aineiston ja mallin välillä odottamattoman suuria jonkin ominaisuuden mukaan (Barlow & Prentice, 1988). Jotta residuaaleja voidaan käyttää, tulee niiden jakauma tietää, kun mallin oletukset ovat voimassa. Tässä luvussa käsitellään martingaali- (Barlow & Prentice, 1988), devianssi- (Therneau, Grambsch & Fleming, 1990) sekä painotettujen Schoenfeldin (Grambsch & Therneau, 1994) residuaalien käyttötapoja.

Martingaaliresiduaalit lasketaan kullekin havaintoyksikölle. Ne saavat arvoja välillä $(-\infty, 1)$ ja summautuvat nollaan. Ykköstä lähellä olevia residuaaleja havaitaan yksilöillä, jotka ovat kuolleet odotettua aikaisemmin, suuret negatiiviset arvot viittaavat ennustetta pidempään elinaikaan. Martingaaliresiduaali voidaan tulkita havaintoyksikön kuolemisen määrän sekä sen odotusarvon erotuksena. Sensuroiduille havainnoille arvo on aina negatiivinen. (Collett, 2003.)

Martingaaliresiduaalien piirtäminen esimerkiksi havaintoyksiköiden indeksien, elinaikojen, elinaikojen järjestyslukujen tai kovariaattien arvojen suhteen voi korostaa tiettyjä yksilöitä, hetkiä tai kovariaattien arvoja, joilla malli ei sovi hyvin aineistoon. Devianssiresiduaalit ovat muunnos martingaaliresiduaaleista. Toisin kuin martingaaliresiduaalit, ne jakautuvat symmetrisesti nollan

suhteen. Devianssiresiduaaleja voi tarkastella samanlaisilla kuvioilla kuin martingaaliresiduaaleja, mutta poikkeavuuksien huomaaminen voi olla helpompaa symmetrisyyden johdosta. (Collett, 2003.)

Schoenfeld (1982) on kehittänyt osittaiset residuaalit, jotka lasketaan jokaiselle havaintoyksikölle kovariaateittain. Grambsch ja Therneau (1994) ovat esitelleet näille residuaaleille painotuksen, joka parantaa residuaalien ominaisuuksia mallin sopivuuden tutkimisessa. Painotetuilla Shoenfeldin residuaaleilla voidaan erityisesti tutkia oletusta suhteellisista riskitiheyksistä. Mikäli oletus ei päde, mallin lineaarinen osa, joka sisältää kertoimet sekä muuttujien arvot, muuttuu ajassa. Painotetut Schoenfeldin residuaalit käyvät kovariaattien kertoimien ajassa muuttuvuuden testaamiseen. Painotettuihin Schoenfeldin residuaaleihin lisätään kovariaateittain kunkin kertoimen estimaatti, ja nämä arvot piirretään elinaikojen funktiona. Mikäli pistejoukko kuvastaa muuta kuin nollakeskistä suoraa, kovariaatin kerroin muuttuu ajassa pistejoukon esittämän funktion mukaisesti. (Collett, 2003.)

Esimerkkinä Coxin mallille kehitetyistä monipuolisista diagnostiikkakeinoista nostan esille graafisen menetelmän, jonka Elja Arjas on esitellyt artikkelissaan 1988. Menetelmällä voidaan tutkia, puuttuuko mallista jokin tärkeä kovariaatti. Sitä voi myös käyttää suhteellisten riskitiheyksien oletuksen tarkasteluun. Kuva piirretään ja tulkitaan seuraavasti: Aineisto ositetaan ryhmiin. Jokaiselle ryhmälle piirretään kuvaaja, jossa vaaka-akselilla on ryhmässä havaittujen kuolemien määrä ja pystyakselilla mallin antama estimaatti kuolleiden määrälle samana hetkenä. Ositteiden kuvaajia verrataan suoraan, jonka kulmakerroin on yksi. Mikäli suorat poikkeavat vertailusuorasta, malli ei ole oikein määritelty tai malliin tulisi lisätä kovariaatti. (Arjas, 1988.)

Riskitiheyskuvio (Karvanen & Harrell, 2009) täydentää esitettyjä diagnostiikkakeinoja. Toisin kuin muut menetelmät, riskitiheyskuvio ei perustu residuaaleihin. Riskitiheyskuvio havainnollistaa yhteen kuvaan mallissa olevien kovariaattien likimääräiset jakaumat sekä mallin antamat suhteelliset riskitiheydet kovariaateittain. Kuvasta voidaan arvioida, ovatko mallin antamat tulokset järkeviä. Riskitiheyskuvio mahdollistaa mallissa olevien kovariaattien keskinäisen merkityksen arvioinnin populaatiotasolla. Riskitiheyskuviota voidaan käyttää myös kovariaatin muunnosten tulkinnalliseen vertailuun. Kuviota käsitellään laajemmin luvuissa 4–5.

3 Ohjelmistot ja aineistot

Tutkielmassa käytetty ohjelmisto on R-ympäristön Windows-versio 3.1.0, joka on nimeltään ”Spring dance” (R Core Team, 2014). Kaikki esitetyt riskitiheyskuviot on piirretty CRANiin julkaistua `rankhazard`-paketin versiota 1.0 käyttäen (Karvanen ja Koski, 2014).

Paketista `rankhazard` löytyvälle riskitiheyskuvion piirtofunktiolle `rankhazardplot` on tehty kattava dokumentaatio (Liite B), jossa julkaistut esimerkit (Liite C) on tehty R-ympäristön `survival`-paketista löytyvillä aineistoilla. Näin funktion käyttäjät pystyvät itse ajamaan esimerkit. Monipuolisuutta pro gradu -työhöni tuovat esimerkit, jotka on tehty Jyväskylän yliopiston gerontologian laitoksella kerätyllä Ikivihreät-aineistolla.

3.1 R-ympäristö

R on ohjelmistokieli sekä ympäristö, joka sisältää suuren määrän erilaisia tilastollisia tapoja tarkastella aineistoa ja sovittaa malleja. R on samankaltainen ohjelmointikielen S kanssa. R on avoimen lähdekoodin ympäristö, mikä mahdollistaa käyttäjälle sen laajentamisen. R-ympäristön saa ilmaiseksi itselleen ja se toimii erilaisilla alustoilla, esimerkiksi Linuxilla, Windowsilla ja MacOS:lla.

Paketti `rankhazard` on laajennus R-ympäristön monipuolisiin grafiikoihin. Funktion `rankhazardplot` toteutuksessa on hyödynnetty R-ympäristössä olevien objektien luokkia sekä S3-järjestelmää, jotka liittyvät läheisesti toisiinsa. Objektin luokka on yksinkertaisimmillaan esimerkiksi numero, merkki, matriisi, lista tai datamatriisi. Luokka voi olla erikoisempi, kuten elinaika. Elinaikaobjektille tulee mm. määrittää elinaika sekä, onko se sensuroitu vai ei.

R-ympäristössä toimii S3-järjestelmä, joka hyödyntää objektin luokkaa. Esimerkiksi `plot`-funktio käyttää S3-järjestelmää: oletuksena se piirtää pisteitä, mutta jos annettavan objektin luokka on aikasarja, se kuvataan automaattisesti viivalla. Luokka ”aikasarja” on R-ympäristössä nimellä `ts`. Tällöin koodi `plot(aikasarjaobjekti)` kutsuu funktiota `plot.ts`, joka säätää piirto-ominaisuudet aikasarjalle.

S3-järjestelmä mahdollistaa sen, että voidaan luoda useita näennäisesti samannimisiä funktioita, jotka toimivat eri tavoin. Nämä funktiot käsittelevät eri luokan objekteja. Kuitenkaan jokaiselle luokalle ei tarvitse tehdä omaa funktiota. Sen varmistaa oletusfunktio, jonka pääte on `.default`. Oletusfunktiota käytetään, mikäli käytettävän objektin luokalle ei ole omaa funktiota.

Jotta `rankhazard`-pakettia voi käyttää, tulee ensin sovittaa Coxin malli. Sovittamiseen voi käyttää joko `survival`- (Therneau, 2014) tai `rms`-pakettia (Harrell, 2014). Ne antavat eri luokkaa olevat Coxin malliobjektit. Paketit ovat käyttäjien tekemiä laajennuksia R-ympäristöön. Seuraavaksi esitellään nämä paketit, ja kuinka niitä on hyödynnetty `rankhazardplot`-funktiossa.

3.1.1 Paketti `survival`

Terry Therneau on luonut `survival`-paketin R-ympäristöön vuonna 2001. Tutkielmassani käytetty versio on 2.37-7, ja se on julkaistu 2014 (Therneau, 2014). Paketissa on funktio `coxph`, jolla voidaan sovittaa erilaisia Coxin malleja. Funktiolla saadut mallit ovat luokaltaan `coxph`. Suurin osa tutkielmassa käytetyistä malleista on sovitettu `coxph`-funktiolla.

Funktion `rankhazardplot.coxph` sisällä käytetään `survival`-paketin funktiota `predict.coxph`. Sillä saadaan suhteellisen riskitiheyden laskemisessa tarvittavat termittäiset ennusteet, jotka on määritelty sivulla 16. Suhteellisen riskitiheyden laskeminen termittäisten ennusteiden pohjalta on käsitelty liitteessä A.4.4.

3.1.2 Paketti `rms`

Toinen tapa sovittaa Coxin malli R-ympäristössä, on käyttää pakettia `rms`. Sen on tehnyt Frank E. Harrell Jr.. Coxin mallin sovittamiseen käytetään funktiota `cph`, ja sen palauttama malli on luokaltaan myös `cph`. Paketti `rms` on julkaistu CRANiin ensimmäistä kertaa vuonna 2009. Käyttämäni paketin versio on 4.2-0 ja se on vuodelta 2014 (Harrell, 2014).

Myös `cph`-funktiolla sovitetut mallit on sisällytetty funktion `rankhazardplot` piirtovalikoimaan. Kun funktiota kutsutaan käyttäen `cph`-mallia, käytetään `rankhazardplot.cph`-funktiota. Sen sisällä käytetään `rms`-paketin funktiota `predict.rms` termittäisten ennusteiden laskemiseen.

3.2 Aineistot `pbc` ja `cgd`

Melkein kaikissa liitteen C esimerkeissä on käytetty `survival` paketista löytyvää `pbc`-aineistoa (Therneau & Grambsch, 2000). Aineistossa on primaarista biliaarista kirroosia eli sappikirroosia (Mustajoki, www.terveyskirjasto.fi, 2013) sairastavia potilaita, joista 312 ovat olleet mukana laajoissa laboratoriomittauksissa. Statusmuuttujaa on muokattu, jotta voidaan sovittaa tavallinen Coxin malli: siirrännäisen saaneet (19/312) on poistettu tarkastelusta. Malleissa käytetyt muuttujat ovat:

time elinaika päivissä

statusbin muokattu statusmuuttuja, 0 = ”havaintoyksikkö sensuroitunut”, 1 = ”havaintoyksikkö kuollut”

albumin seerumi albumiiniarvo (g/dl)

age ikä vuosissa

ast erään entsyymin (engl. aspartate aminotransferase) pitoisuus veressä (U/ml)

bili seerumi bilirubiiniarvo (mg/dl)

copper kuparimäärä virtsassa (μg/päivä)

edema 0 = ”ei turvotusta”, 0.5 = ”onnistuneesti hoidettu tai hoitamaton turvotus”, 1 = ”turvotus hoidosta huolimatta”

protime standardisoitu veren hyytymisaika

sex sukupuoli, m = ”mies”, f = ”nainen”

stage maksan kudoksen muutoksen aste, 1 = ”kolangiitti tai periportaalinen hepatiitti”, 2 = ”periportaalinen fibroosi tai hepatiitti”, 3 = ”septaalinen fibroosi”, 4 = ”kirroosi” (Locke, Therneau, Ludwig, Dickson & Lindor, 1996; suomennokset Färkkilä, 1993)

Esimerkki C.12 (sivu 95) on sovitettu käyttäen `cgd`-aineistoa (Fleming & Harrington, 1991). Aineisto on saatavilla `survival`-paketissa. Se on valittu esimerkiksi aineistosta, jossa kustakin havaintoyksiköstä voi olla useita rivejä havaintomatriisissa. Havaintoyksiköt sairastavat kroonista sairautta, ja tutkimuksen aikana on mitattu aikaa infektioon. Jokaiselle infektiolle on oma aikaväli, joka alkaa edellisestä sairastumisesta ja päättyy seuraavaan. Havaintoyksikköjä on aineistossa 128. Heistä 84 ei ole sairastunut kertaakaan tutkimuksen aikana, 27 on sairastunut kerran, 9 on sairastunut kaksi kertaa ja 5 on sairastunut kolme kertaa. Lisäksi neljän, viiden ja seitsemän sairastumisen kertoja on havaittu, kutakin yhdellä havaintoyksiköllä. Seuranta on kestänyt eri havaintoyksiköillä 91–439 päivää. Mallissa käytetyt muuttujat ovat:

tstart aikavälin alku

tstop aikavälin loppu

status status aikavälin lopussa, 1 = ”sairastunut”, 0 = ”sensuroitunut”

height pituus tutkimuksen alussa (cm)

id havaintoyksikön identifioiva numero

steroids käyttääkö steroideja tutkimuksen alussa, 1 = ”kyllä”, 0 = ”ei”

treatment käsittely, ”placebo” tai ”rIFN-g”

3.3 Ikivihreät-aineisto

Gerontologian tutkimuskeskuksella on jo 1980-luvulta ollut käynnissä projekti Ikivihreät, joka sisältää useita seurantatutkimuksia. Aineistot on kerätty, jotta voidaan tutkia iäkkäiden ihmisten psyykkistä ja fyysistä toimintakykyä sekä niiden muutosta ikääntymisen myötä. Pro gradu -työssäni on käytössä aineisto,

jossa on laboratoriomittauksia 295:ltä vuoden 1989 aikana 75 vuotta täyttäneeltä jyväskyläläiseltä. Mittauksia on toistettu seurannan aikana viiden vuoden välein. (Gerontologian tutkimuskeskus, Evergreen-project ja Ikivihreät-projekti.)

Aineistossa on 104 miestä ja 191 naista. Havainnoista vain 8 on sensuroituja, sillä kaikkien kuolleiden kuolinpäivä on tiedossa. Elinaikaa on mitattu 12.9.2012 asti. Tällöin tutkittavat henkilöt ovat olleet noin 98-vuotiaita.

Aineisto on tallennettu nimelle `gero`. Malleissa käytettävät muuttujat ovat:

eloaika Havaintoyksikön elinaika päivissä 1.1.1990 eteenpäin.

delta Mitatun elinajan status. Arvot 0 = ”sensuroitu”, 1 = ”kuollut”.

sp Sukupuolifaktori, jonka tasot ovat ”nainen” ja ”mies”. Vertailutaso on ”nainen”.

bmi75, bmi80 Painoindeksi (engl. *body mass index*) 75 ja 80 vuoden iässä.

nopeus75, nopeus80 Kävelynopeus (metriä sekunnissa) 10 metrin matkalla 75 ja 80 vuoden iässä.

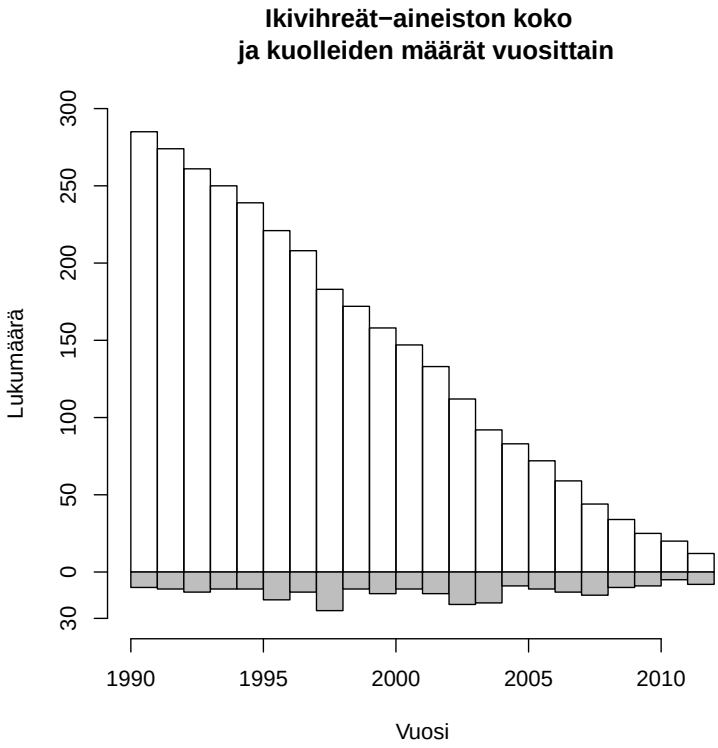
vaik75, vaik80 Päivittäisissä toiminnoissa koettujen vaikeuksien pistemäärä 75 ja 80 vuoden iässä.

tervtila Terveydentilaa kuvaava faktori, jonka tasot 1 = ”epätavallisen hyvä” (8 kpl), 2 = ”hyvä” (144 kpl), 3 = ”vähemmän hyvä, huono” (135 kpl), 4 = ”erittäin huono” (7 kpl). Vertailutaso on 1. Puuttuvia havaintoja on yksi.

Taulukon 1 on tunnusluvuista huomataan, että painoindeksin jakauma säilyy likimain samana ikävuosien 75 ja 80 välillä, kuten myös päivittäisissä toiminnoissa koettujen vaikeuksien pistemäärän jakauma. Suurin muutos tapahtuu kävelynopeudessa, joka hidastuu. Tunnusluvut on laskettu kaikista havainnoista. Kuvaan 1 on kuvattu elinaikojen histogrammit kahdesta näkökulmasta: elossa olevien sekä vuoden aikana kuolleiden lukumäärät.

Taulukko 1: Ikivihreät-aineiston jatkuvien muuttujien tunnuslukuja.

Muuttuja	Minimi	Q_1	Mediaani	Keskiarvo	Q_3	Maksimi	NA
bmi75	17.96	24.22	26.78	27.08	29.34	44.80	0
bmi80	17.04	24.85	26.94	26.98	29.14	39.53	107
nopeus75	0.465	1.351	1.587	1.589	1.852	3.125	2
nopeus80	0.322	1.099	1.389	1.380	1.613	2.778	112
vaik75	10	11	15	16.06	20	41	10
vaik80	10	11.25	14	17.04	21	46	85



Kuva 1: Ikivihreät-aineistossa olevien henkilöiden kuolinpäivien histogrammi harmaalla, sekä vuoden lopussa elossa olevien histogrammi valkoisella. Kuvasta puuttuvat viimeisen vuoden tiedot: vuoden 2012 aikana 12.9.2012 mennessä jäljellä olevista 12 henkilöstä kuoli 4, joten tutkimusajan lopussa elossa oli vielä 8 henkilöä.

4 Riskitiheyskuvio

Tässä luvussa esitetään suhteellisen riskitiheyden laskeminen sekä riskitiheyskuvion piirtäminen. Sen jälkeen käydään läpi riskitiheyskuvion tulkintaa sekä luottamusvälien laskeminen ja piirtäminen suhteelliselle riskitiheydelle. Karvanen ja Harrell (2009) korostavat, että riskitiheyskuviota käytetään ennen kaikkea havainnollistamistarkoitukseen. Raportoitaessa käytetään alkuperäisiä Coxin mallin antamia riskitiheyssuhteita. Täysimittaista data-analyysiä tehtäessä sovitettua mallia kannattaa tarkastella useita diagnostiikkakeinoja käyttäen. Riskitiheyskuvion tulkitsemisen lisäksi on tällöin tarpeellista tehdä esimerkiksi luvuissa 2.4 ja 5 esitettyjä residuaalitarkasteluja.

Karvanen ja Harrell (2009) toteavat, että riskitiheyskuviota voidaan käyttää myös muiden elinaika- ja regressiomallien havainnollistamiseen. Olen ideoinut tavan soveltaa riskitiheyskuviota ajassa muuttuvien kovariaattien malliin, mikä on esitelty luvussa 4.5.

Tutkielmassa esitetyt riskitiheyskuviot on piirretty R-ympäristön `rankhazard`-pakettia käyttäen (Karvanen & Koski, 2014). Kaikkiin kuviin annetaan piirtokoodi, jotta lukija näkee useita esimerkkejä riskitiheyskuvion piirtämisestä. Koodien lukemista helpottaa, mikäli ensin tutustuu paketin ja piirtofunktion toimintaan. Niitä käsitellään luvussa 6 sekä liitteessä A.

4.1 Suhteellisen riskitiheyden laskeminen ja käsitteitä

Kovariaattiin j liittyvän Coxin mallin kertoimen β_j avulla saadaan kovariaatille j riskitiheyssuhde, joka on $\exp(\beta_j)$. Se kertoo, monikertaiseksi riskitiheys muuttuu, mikäli kovariaatin j arvo kasvaa yhden yksikön, ja muiden mallissa olevien kovariaattien arvot pysyvät samana. Riskitiheyssuhde kuvaa siis kovariaatin merkitystä riskitiheyteen. Kaikki muuttujat eivät ole kuitenkaan keskenään vertailukelpoisia pelkän riskitiheyssuhteen perusteella, sillä kovariaattien vaihteluvälit voivat olla hyvin erilaiset. Jopa mittayksikön vaihtaminen esimerkiksi kilogrammoista grammoihin muuttaa riskitiheyssuhdetta.

Mallin mukaan voidaan myös estimoida riskitiheysfunktio $h_i(t)$ kullekin henkilölle i. Kutsutaan suhteelliseksi riskitiheydeksi riskitiheyttä, joka on skaalattu toisella riskitiheydellä. Tätä skaalaavaa riskitiheyttä kutsutaan vertailuriskitiheydeksi ja merkitään $h_{\text{vert}}(t)$. Suhteellinen riskitiheys henkilölle i on siis

$$\frac{h_i(t)}{h_{\text{vert}}(t)}.$$

Suhteellisen riskitiheyden jakaumaa käyttäen voidaan vertailla eri kovariaattien merkitystä riskitiheyteen. Suhteellisen riskitiheyden jakaumaa estimoitaessa sen arvo lasketaan jokaiselle henkilölle ja kovariaatille erikseen. Riskitiheys henkilölle i noudattaa Coxin mallin riskitiheysfunktiota, joka on määritelty kaavassa (1) sivulla 6. Vertailuriskitiheys määritellään riskitiheytenä kovariaatin vertailukohdassa $x_{\text{ref},j}$. Vertailukohtana voidaan käyttää esimer-

kiksi kovariaatin mediaania, tai käytettävä arvo voidaan valita sisällöllisten kysymysten pohjalta.

Henkilön i kovariaatin j arvoon $x_{i,j}$ liittyvä suhteellinen riskitiheys saadaan siis jakamalla $h_0(t)\exp(\beta_j x_{i,j})$ vertailuriskitiheydellä $h_0(t)\exp(\beta_j x_{\mathrm{ref},j})$. Tulkittaessa oletetaan muiden kuin kovariaatin j arvojen olevan samoja käytetyissä riskitiheyksissä, jotta niihin liittyvät termit supistuvat laskussa. Coxin mallin tapauksessa kovariaattiin j liittyvä suhteellinen riskitiheys otoksessa olevalle henkilölle i on

$$\frac{h_0(t)\exp(\beta_j x_{i,j})}{h_0(t)\exp(\beta_j x_{\mathrm{ref},j})} = \exp(\beta_j x_{i,j} - \beta_j x_{\mathrm{ref},j}) \tag{3}$$

$$= \exp\big(\beta_j (x_{i,j} - x_{\mathrm{ref},j})\big) \tag{4}$$

$$= \big(\exp(\beta_j)\big)^{(x_{i,j} - x_{\mathrm{ref},j})}. \tag{5}$$

Suhteellinen riskitiheys saa arvoja nollasta äärettömään. Se on vertailukohdassa yksi. Mikäli kovariaatin j arvojoukko sekä kerroin β_j on positiivinen, suhteellinen riskitiheys on vertailukohtaa pienemmillä kovariaatin arvoilla alle yksi ja suuremmilla arvoilla yli yksi. Mikäli kerroin on negatiivinen, suhteellinen riskitiheys on ennen vertailukohtaa suurempi kuin yksi ja vertailukohdan jälkeen nollan ja yhden välissä.

Kaavassa (4) on sievin muoto suhteellisesta riskitiheydestä. Eri muodot on esitetty, koska niistä nähdään huomionarvoisia asioita. Kaava (5) näyttää, että kovariaatin j suhteellinen riskitiheys henkilölle i voidaan laskea korottamalla kovariaatin j riskitiheyssuhde potenssiin $x_{i,j} - x_{\mathrm{ref},j}$.

Kaavassa (3) on kovariaattiin j liittyvä termi Coxin mallin lineaarisesta osasta $\beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_p x_{i,p}$. Ennuste lineaariselle osalle saadaan korvaamalla kertoimet β_j niiden estimaateilla $\hat{\beta}_j$. Ennuste voidaan laskea erikseen jokaiselle lineaarisessa osassa olevalle termille, ja koko ennuste saadaan niiden summana. Kovariaatille j lasketaan vertailuarvo $\hat{\beta}_j x_{\mathrm{ref},j}$, joka on termittäinen ennuste kovariaatin vertailukohdassa. Tämä vertailuarvo vähennetään kyseisen kovariaatin termittäisestä ennusteesta henkilölle i.

Tärkeää laskutavassa on huomata, että termittäiset ennusteet $\hat{\beta}_j x_{i,j}$ on tässä esitetty jatkuvalle muuttujalle, jolle estimoidaan mallissa yksi kerroin. Tällaisia ovat aineistossa olevat muuttujat sekä niiden yksinkertaiset muunnokset. Yksinkertainen muunnos on esimerkiksi logaritmi alkuperäisistä arvoista. Tällöin $x_{i,j}$ sisältää muunnetun arvon. Mikäli jatkuvalle muuttujalle on sovitettu monimutkainen muunnos, mikä tarkoittaa useita kertoimia yhdelle muuttujalle, suhteellinen riskitiheys saadaan laskettua käyttämällä termin $\hat{\beta}_j x_{i,j}$ tilalla ennustetta arvolle $x_{i,j}$ ja vastaavasti $\hat{\beta}_j x_{\mathrm{ref},j}$ tilalla ennustetta vertailukohdalle. Kyseisen ennusteen laskeminen voi olla monimutkaisempaa kuin estimoidun kertoimen ja kovariaatin arvon tulo.

Kaavassa (6) on esitetty kaava (3) kahdelle kovariaatille A ja B, kun perusriskitiheys on jo supistettu pois. Potenssien laskusääntöjä käyttämällä saadaan

Taulukko 2: Tutkielmassa käytetyt suhteelliseen riskitiheyteen liittyvät käsitteet suomeksi. Englanninkieliset termit ovat pääosin peräisin artikkelista *Visualizing covariates in proportional hazards model* (Karvanen & Harrell, 2009) tai `rankhazardplot`-funktion dokumentaatiosta (liite B). Lisäksi havainnollistavat kaavat. Indeksi i kuvaa havaintoyksikköä ja indeksi j kovariaattia.

suomi	englanti	kaava/arvo
suhteellisen riskitiheyden jakauma	distribution of the relative hazard	-
riskiheyskuvio	rank-hazard plot	-
vertailukohta	reference point	$x_{\text{ref},j}$
vertailuarvo	reference value	$\beta_j x_{\text{ref},j}$
vertailuriskitiheys	reference hazard	$h_0(t)\exp(\beta_j x_{\text{ref},j})$
riskitiheyssuhde	hazard ratio	$\exp(\beta_j)$
suhteellinen riskitiheys	relative hazard	$\exp\left(\beta_j(x_{i,j} - x_{\text{ref},j})\right)$
suhteellisen riskitiheyden logaritmi	logarithm of the relative hazard	$\beta_j(x_{i,j} - x_{\text{ref},j})$
referenssi eli suhteellisen riskitiheyden arvo vertailukohdassa	reference	1
referenssi (suhteellisen riskitiheyden logaritmille)	reference (logarithmic)	0

kahden kovariaatin suhteellinen riskitiheys tulomuotoon (7):

$$\frac{\exp(\beta_A x_{i,A} + \beta_B x_{i,B})}{\exp\left(\beta_A x_{\text{ref},A} + \beta_B x_{\text{ref},B}\right)} = \exp\left(\beta_A(x_{i,A} - x_{\text{ref},A}) + \beta_B(x_{i,B} - x_{\text{ref},B})\right) \quad (6)$$

$$= \exp\left(\beta_A(x_{i,A} - x_{\text{ref},A})\exp\left(\beta_B(x_{i,B} - x_{\text{ref},B})\right)\right). \quad (7)$$

Mikäli halutaan laskea vielä useamman kovariaatin yhteinen suhteellinen riskitiheys, kaava yleistyy vastaavasti.

Suhteellisen riskitiheyden laskutavasta johtuen sen arvo vertailukohdassa on 1. Suhteellisen riskitiheyden sijaan voidaan käyttää myös suhteellisen riskitiheyden logaritmin arvoa $\beta_j(x_{i,j} - x_{\text{ref},j})$. Tällöin arvo vertailukohdassa on 0. Nimike, jolla viitataan tähän vertailukohdassa havaittavaan arvoon, on referenssi. Se on 1 tai 0 riippuen siitä, puhutaanko suhteellisesta riskitiheydestä vai sen logaritmista.

Taulukkoon 2 on koottu suhteelliseen riskitiheyteen liittyvät käsitteet sekä suomeksi, englanniksi että kaavoina. Suomennokset on luotu tätä tutkielmaa varten.

4.1.1 Suhteellinen riskitiheys faktorille

Faktori on esimerkki muuttujista, joille estimoidaan Coxin mallissa useita kertoimia. Faktoreiden tapauksessa, jokaiselle muuttujan tasolle paitsi vertailutasolle tarvitaan oma kerroin. Muiden faktorin tasojen vaikutusta riskitiheyteen verrataan tähän vertailutasoon. Mikäli kerroin on positiivinen, riskitiheys on suurempi kuin vertailutasolla. Jos taas kerroin on negatiivinen, riskitiheys on pienempi kuin vertailutasolla. Mikäli tilastollinen testi ei osoita kertoimen eroavan nollasta, riskitiheys ei eroa tilastollisesti merkitsevästi vertailutasosta.

Olkoon k faktorin A taso henkilöllä i ja β_{A_k} kyseisen tason kerroin. Vastaavasti $\beta_{A_{\text{ref}}}$ on kyseisen faktorin vertailukohdan kerroin, joka on nolla. Nyt suhteellisen riskitiheyden kaava (3) saa muodon

$$\frac{h_0(t)\exp(\beta_{A_k})}{h_0(t)\exp(\beta_{A_{\text{ref}}})} = \exp(\beta_{A_k} - \beta_{A_{\text{ref}}}) = \exp(\beta_{A_k}),$$

joka on kyseisen tason riskitiheyssuhde. Faktorin tapauksessa tasojen riskitiheyssuhteet sisältävät informaation kovariaatin suhteellisesta riskitiheydestä.

Malleissa käytettävät faktorit voivat olla myös luokitteluasteikollisia, eikä faktorin suhteellisen riskitiheyden kuvaaja ole välttämättä monotoninen.

4.2 Suhteellisen riskitiheyden jakauman kuvaaja

Käydään seuraavaksi läpi suhteellisen riskitiheyden jakauman kuvaajan piirtämisen periaatteet. Luvussa 4.3 perehdytään kuvan käyttämiseen mallin tulkinnassa.

Riskitiheyskuviossa pystyakselilla on suhteellinen riskitiheys tai sen logaritmi. Suhteellisen riskitiheyden asteikkona käytetään logaritmista asteikkoa, joka on suhteellinen, ei lineaarinen. Suhteellisuus tarkoittaa sitä, että esimerkiksi välimatka 0.5–1 piirretään yhtä pitkäksi kuin välillä 1–2, sillä molemmilla väleillä muutos on kaksinkertainen. Siten myös välimatka arvojen 5 ja 10 välillä on yhtä pitkä kuin edellä mainittujen. Mikäli kuvataan suhteellisen riskitiheyden logaritmi, asteikko on lineaarinen. Kuvaajat näyttävät samalta, piirretään se suhteellisena riskitiheytenä tai sen logaritmina. Ainoastaan asteikko muuttuu.

Samaan riskitiheyskuvioon voidaan piirtää eri kovariaattien suhteellisen riskitiheyden jakaumia. Koska kovariaattien arvojen vaihteluväli voi olla hyvin erilainen kovariaattien välillä, vaaka-akselin koordinaattina käytetään skaalattuja järjestyslukuja. Tällöin kovariaattien arvot siirtyvät tasavälisesti välille $[0, 1]$. Skaalatut järjestysluvut mahdollistavat vertailun jatkuvien muuttujien ja faktorien välillä. Lisäksi kuvaaja estimoi kovariaatin jakaumaa perusjoukossa ja suhteellisen riskitiheyden jakaumat ovat keskenään vertailukelpoisia. (Karvanen & Harrell, 2009.)

Skaalattujen järjestyslukujen määrittämisen voi esittää kaavan avulla. Olkoon $r_{i,j}$ järjestysluku henkilön i kovariaatin j arvolle suuruusjärjestyksessä,

Taulukko 3: Coxin mallin kertoimet, riskitiheyssuhteet ja merkitsevyyden testaus Ikivihreät-aineistoon sovitetulle mallille numero 1, jossa on selittävänä muuttujana terveydentila-faktori, n = 294.

Muuttuja	$\hat{\beta}$	$\exp(\hat{\beta})$	$\text{se}(\hat{\beta})$	p
tervtila_2	0.475	1.61	0.388	0.220
tervtila_3	0.880	2.41	0.389	0.024
tervtila_4	2.026	7.59	0.542	<0.001

ja n on havaintojen määrä. Tällöin skaalattu järjestysluku kyseisen henkilön kovariaatille on

$$\frac{r_{i,j} - 1}{n - 1}.$$

Ensimmäisen havainnon skaalattu järjestysluku on 0 ja viimeisen 1.

Kullekin havaintoyksikölle lasketaan kovariaateittain suhteellinen riskitiheys kovariaatin alkuperäistä arvoa käyttäen sekä skaalattu järjestysluku. Jälkimmäinen määrää vaaka-akselin koordinaatin ja ensimmäinen pystyakselin koordinaatin. Suhteellisen riskitiheyden jakauma piirretään yhdistämällä nämä pisteet viivalla vasemmalta oikealle. Kuviosta nähdään suhteellisen riskitiheyden vaihteluväli kullakin kovariaatilla sekä se, kuinka suuri osa havainnoista kuuluu ryhmään, jossa suhteellinen riskitiheys on suuri tai pieni. Kuvan tulkintaa voi helpottaa referenssisuora, joka on vaakasuora referenssin kohdalla.

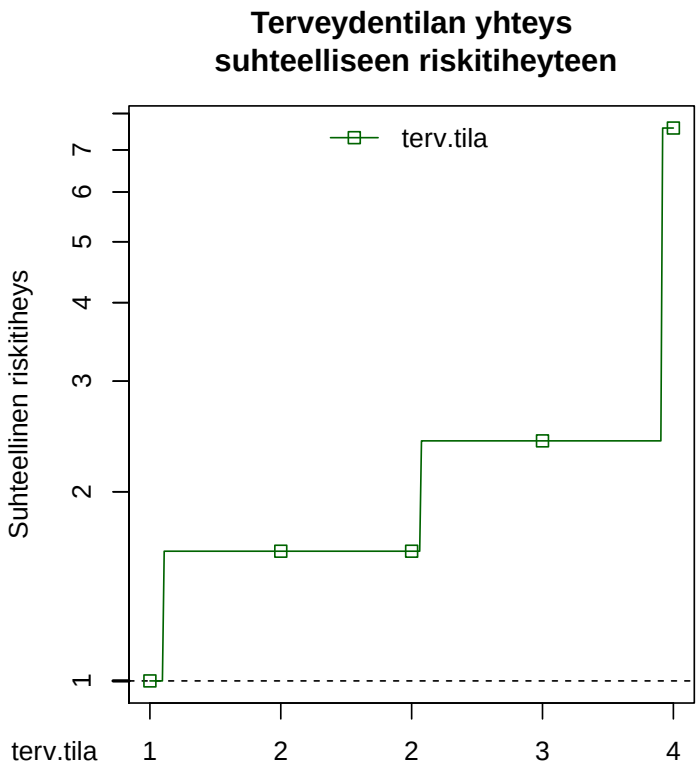
Kun piirtämisessä käytetään järjestyslukuja, tieto kovariaatin arvoista häviää. Siksi suhteellisen riskitiheyden jakauman kuvaan merkitään vaaka-akselille kunkin kovariaatin minimi, alakvartiili, mediaani, yläkvartiili ja maksimi vastaavasti kohtaan $0, 0.25, 0.5, 0.75$ ja 1. Näin kuvio havainnollistaa samalla aineistoa, sillä siitä nähdään likimain kovariaattien jakaumat.

Otetaan yksinkertainen esimerkki, jossa riskitiheyttä selitetään yhdellä faktorilla, terveydentilalla. Terveydentila on neliluokkainen faktori, jonka tasot 1 = ”epätavallisen hyvä”, 2 = ”hyvä”, 3 = ”vähemmän hyvä, huono”, 4 = ”erittäin huono”. Vertailutaso on 1. Malli on sovitettu koodilla:

```
aika <- Surv(gero$eloaika, gero$delta)
malli1 <- coxph(aika ~ as.factor(tervtila), data = gero,
  x = TRUE)
```

Kuvaan 2 on piirretty terveydentilan suhteellisen riskitiheyden jakauma seuraavalla koodilla:

```
par(mar = c(2, 5, 4, 2) + 0.1);par(mfrow = c(1, 1))
rankhazardplot(malli1, data = gero, legendtext = "terv.tila",
  col = "darkgreen", refline = TRUE,
  main= "Terveydentilan yhteys suhteelliseen riskitiheyteen",
  ylab = "Suhteellinen riskitiheys")
```



Kuva 2: Riskitiheyskuvio Coxin mallista, jossa riskitiheyttä selitetään terveydentilalla. Mitä suuremman arvon terveydentila-muuttuja saa, sitä huonompi terveydentila henkilöllä on. Vertailutasona on epätavallisen hyvä terveydentila.

Kuvasta nähdään eri tasojen riskitiheyssuhteet, kun vertailutasona on epätavallisen hyvä terveydentila. Samat riskitiheyssuhteet on esitetty taulukossa 3. Riskitiheyskuviosta nähdään lisäksi terveydentilan jakauma. Ensimmäinen ja viimeinen tasanne kuvaajassa ovat kapeita, toinen ja kolmas leveitä. Tasanteet vastaavat terveydentilan tasoja 1–4, minkä näkee vaaka-akselin asteikolta. Siis luokkiin 2 ja 3 kuuluvia henkilöitä on suunnilleen yhtä paljon kuten myös luokkiin 1 ja 4 kuuluvia.

Luokat 2 ja 3 sisältävät arviolta 95 prosenttia havaintoyksiköistä ja loput 5 prosenttia jakautuvat melko tasan luokkien 1 ja 4 välille. Sivulla 13 esitetyn terveydentila-muuttujan jakauman perusteella saadaan laskettua vastaavat tarkat arvot, joiden mukaan esimerkiksi luokkiin 1 ja 4 kuuluvien osuus on $15/294 \approx 0.051$ eli 5.1 prosenttia. Kunnoltaan erittäin huonojen henkilöiden riskitiheyssuhde verrattuna kunnoltaan epätavallisen hyviin on liki 7.6, mutta tähän riskiryhmään kuuluvia henkilöitä on todella vähän. Koska ryhmä 1 on vertailuryhmä, sen suhteellinen riskitiheys on 1.

Hieman yli puolet havaintoyksiköistä kuuluu joko epätavallisen hyvään tai hyvään kuntoluokkaan. Tämä nähdään siitä, että vaaka-akselille kirjattu mediaani on 2, ja kuvaaja nousee heti mediaanin jälkeen. Mediaani on kuvassa

kohdassa 0.5 ja suhteellisen riskitiheyden arvo on 1.61, mikä on hyvän kunnon riskitiheyssuhde.

Kuvan pystyakseli on logaritminen, ja se on kuvattu välillä 1–8. Asteikkomerkit on piirretty numeerisen erotuksen mukaan tasavälein, jotta arvot voi päätellä niillekin asteikkomerkeille, joille luku ei mahdu tulostumaan kuvaan.

Mikäli kovariaatti on mallissa lineaarisena, suhteellisen riskitiheyden kuvaajan muodosta voi päätellä kovariaatin jakaumaa seuraavasti: Mikäli kovariaatin jakauma on tasajakauma, suhteellisen riskitiheyden kuvaaja on suora viiva. Vaaka-akselin suuntainen kuvaaja kertoo siitä, että kovariaatin arvot eivät muutu, eli kyseisiä arvoja on havaittu paljon. Jyrkkä nousu kuvaajassa taas osoittaa, että kovariaatin arvoissa tapahtuu suuri äkillinen muutos. Tämä näkyy myös faktoreiden suhteellisen riskitiheyden jakauman kuvaajassa, joka on porrasfunktio.

4.3 Riskitiheyskuvion tulkinta

Riskitiheyskuvio näyttää nopean yleiskatsauksen mallissa olevien kovariaattien keskinäisestä merkityksestä kuolemisen riskiin. Satunnaisotoksen tapauksessa suhteellisen riskitiheyden jakauma estimoi jakaumaa populaatiossa. Tällöin kuvion perusteella voidaan tehdä päätelmiä populaatiotasolla. Jos taas kyseessä on esimerkiksi tapaus-verrokkitutkimus, kuviota voi käyttää havainnollistamaan aineistoa ja tuloksia, mutta sen perusteella ei voi tehdä päätelmiä koko populaatiosta. (Karvanen & Harrell, 2009.)

Kiinnostava kysymys voi olla esimerkiksi ”liittyykö tupakointiin vakavampi riski sepelvaltimotautiin kuin ylipainoon”. Riski ei muodostu pelkästään Coxin mallin kertoimista vaan myös näiden kovariaattien ilmenemisestä populaatiossa. Jos esimerkiksi populaatiossa on enemmän ylipainoisia kuin tupakoitsijoita, voidaan sepelvaltimotaudin esiintymistä populaatiotasolla vähentää vaikuttamalla ensisijaisesti ylipainoisiin. Riskitiheyskuviota käytettäessä tulee huomata, että kaikki mallissa olevat kovariaatit vaikuttavat estimoituihin kertoimiin ja sitä kautta suhteelliseen riskitiheyteen. (Karvanen & Harrell, 2009.)

Ajatellaan esimerkiksi mallintavamme sepelvaltimotaudin riskiä Coxin suhteellisten riskitiheyksien mallilla. Olkoon estimoitu riskitiheyssuhde tupakoinnille 2 (1 = ”tupakoi”, 0 = ”ei tupakoi”) ja painoindeksille 1.1. Käytetään vertailukohtana tupakoimatonta henkilöä, jonka painoindeksi on 25. Lasketaan suhteellinen riskitiheys kaavaa (5) käyttäen. Faktorille suhteellinen riskitiheys muodostuu tasojen riskitiheyssuhteista, eli se on tupakoimattomille 1 ja tupakoiville 2. Kaavalla (7) saadaan kahden kovariaatin yhteinen suhteellinen riskitiheys. Painoindeksiltään 25 olevan tupakoitsijan suhteellinen riskitiheys on $2 \cdot 1.1^{25-25} = 2$. Lähes sama suhteellinen riskitiheys saadaan tupakoimattomalle henkilölle, jonka on painoindeksiltään 33 eli $1 \cdot 1.1^{33-25} \approx 2.1$. Siis kahdeksan yksikön nousu painoindeksissä tupakoimattomalle henkilölle vastaa samanlaista riskiä, joka liittyy tupakointiin.

Suhteellisen riskitiheyden arvo riippuu valitusta vertailukohdasta. Jotta

Taulukko 4: Coxin mallin kertoimet, riskitiheyssuhteet ja merkitsevyyden testaus Ikivihreät-aineistoon sovitetulle mallille numero 2, jossa selittävänä muuttujana sukupuoli, painoindeksi, kävelynopeus ja päivittäisissä toimissa koettujen vaikeuksien pistemäärä 75 vuoden iässä. Sukupuolen vertailutaso on nainen. Otoskoko on 283.

Muuttuja	$\hat{\beta}$	$\exp(\hat{\beta})$	$\text{se}(\hat{\beta})$	p
sukupuoli_mies	0.572	1.77	0.139	<0.001
painoindeksi	-0.071	0.93	0.017	<0.001
nopeus	-0.952	0.39	0.184	<0.001
vaikeudet	0.054	1.06	0.012	<0.001

suhteelliset riskitiheydet eri kovariaateilla ovat vertailukelpoisia, kannattaa vertailukohdat valita harkiten. Mikäli ei löydy sisällöllisesti perusteltua syytä tietylle vertailukohdalle, on suositeltavaa käyttää mediaaneja. Sisällöllinen peruste voi olla esimerkiksi ”normaalitilan” käyttäminen vertailukohtana. Esimerkiksi ”normaalitilan” raja painoindeksillä voisi olla normaalipainon yläraja 25 kg/m^2.

Otetaan seuraavaksi esimerkki, jossa mallissa on useita selittäviä muuttujia. Ikivihreät-aineistoon on sovitettu Coxin malli, jossa riskitiheyttä selitetään sukupuolella, painoindeksillä, kävelynopeudella sekä päivittäisissä toiminnoissa koettujen vaikeuksien pistemäärällä 75 vuoden iässä. Sukupuoli on faktori, jonka vertailutaso on nainen. Tähän malliin viitataan jatkossa mallina 2. Malli 2 on sovitettu koodilla:

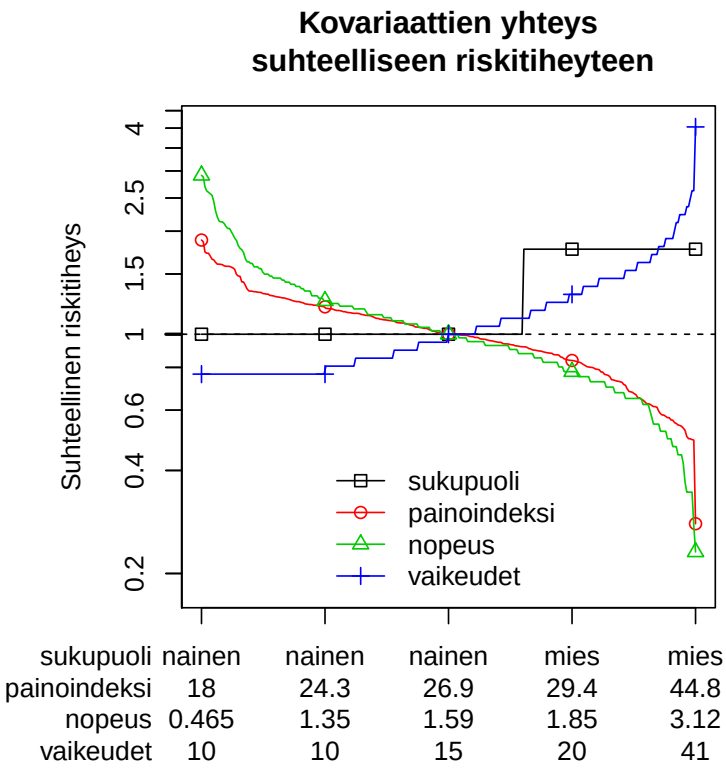
```
aika <- Surv(gero$eloaika, gero$delta)
malli2 <- coxph(aika ~ sp + bmi75 + nopeus75 + vaik75,
  data = gero, x = TRUE)
```

Mallin kertoimet ovat taulukossa 4.

Kuvaan 3 on piirretty riskitiheyskuvio mallissa olevista kovariaateista seuraavalla koodilla:

```
par(mar = c(5, 6, 4, 2) + 0.1);par(mfrow = c(1, 1))
rankhazardplot(malli2, data = gero, ylim = c(0.18, 4.1),
  legendlocation = "bottom", refline = TRUE,
  main = "Kovariaattien yhteys suhteelliseen riskitiheyteen",
  ylab = "Suhteellinen riskitiheys", legendtext = c("sukupuoli",
  "painoindeksi", "nopeus", "vaikeudet"))
```

Vaaka-akselilla ovat kovariaattien minimi, kvartiilit sekä maksimi. Kuvaan on piirretty vain täydelliset, mallin sovituksessa mukana olleet havainnot, joten havaitut tunnusluvut eivät ole täsmälleen samat kuin taulukossa 1 on esitetty.



Kuva 3: Riskitiheyskuvio Coxin mallista, jossa riskitiheyttä selitetään sukupuolella, painoindeksillä, kävelynopeudella sekä päivittäisissä toiminnoissa koettujen vaikeuksien pistemäärällä. Sukupuolen vertailutaso on nainen, muiden muuttujien vertailukohtana on käytetty mediaania.

Sukupuoli-faktori on poikkeus, sillä tasoilla ”nainen” ja ”mies” ei ole järjestystä, joten ei ole kyseisiä tunnuslukujakaan. Nyt järjestys määräytyy sen mukaan, että ensin on vertailutaso. Pystyakselin asteikko on logaritminen, sillä kuvaan on piirretty suhteelliset riskitiheydet. Jotta asteikko on luettavissa, numeerinen välitys muuttuu referenssin kohdalla. Referenssi on korostettu kuvaan mustalla katkoviivalla, joka menee pitkälti päällekkäin sukupuoli-muuttujan suhteellisen riskitiheyden jakauman kuvaajan kanssa. Ennen referenssiä, joka on 1, asteikkomerkit ovat 0.2 välein ja referenssin jälkeen 0.5 välein. Siis ensimmäinen selitteetön asteikkomerkki alhaalta päin vastaa arvoa 0.8 ja toinen arvoa 2.

Riskitiheyskuviosta nähdään, että kävelynopeus ja painoindeksi ovat jatkuvasti mitattuja muuttujia, sillä niiden kuvaajat ovat melko sileitä. Tällöin on havaittu paljon erilaisia arvoja. Vaikeudet-muuttuja saa vain kokonaislukuarvoja, ja sen kuvaajassa on nähtävissä portaita, kun arvot kasvavat yhden yksikön. Asteikon loppupuolella ei havaita kaikkia arvoja ja samoja arvoja ilmenee vähemmän. Tästä syystä porrasmainen muoto silottuu maksimia kohden. Maksimihavainto näyttää myös olevan melko poikkeava sitä edeltävään ver-

rattuna, sillä kuvaaja kääntyy pystysuoraksi. Sama huomataan painoindeksin maksimissa. Sukupuoli saa vain kahta arvoa, ja kuvaajaan tulee yksi porras.

Kuvaajille on piirretty pisteet minimiin, kvartiileihin sekä maksimiin. Pisteet auttavat suhteellisen riskitiheyden lukemisessa noissa arvoissa. Kuvaajasta ei voi nähdä tiettyyn havaintoyksikköön liittyviä arvoja, vaan kaikkia kovariaatteja käsitellään erikseen.

Taulukosta 4 nähdään, että painoindeksillä sekä kävelynopeudella on negatiivinen kerroin: mitä suurempi kävelynopeus tai painoindeksi havaintoyksiköllä on, sitä pienempi on kuolemisen riskitiheys. Tällöin suhteellisen riskitiheyden kuvaaja on laskeva. Positiivinen kerroin vastaa nousevaa suoraa. Luvussa 5.2 tutkitaan, onko painoindeksiä sopivaa mallintaa lineaarisena muuttujana.

Nopeuden kerroin on yli 13-kertainen painoindeksin kertoimeen nähden. Kuitenkin kuvaajat ovat erittäin samankaltaiset. Tämä johtuu siitä, että nopeus saa paljon pienempiä arvoja kuin painoindeksi. Kuvaajat alakvartiilista maksimiin kulkevat lähes päällekkäin. Pienimmillä arvoilla käyrät eroavat siten, että nopeuden suhteellinen riskitiheys on suurempi kuin painoindeksin. Minimissä arvo nopeudelle on noin 3 ja painoindeksille alle 2.

Vertailtaessa kovariaattien suhteellisia riskitiheyksiä keskenään, huomataan, että vaikeudet-muuttujaan liittyvä kuvaaja nousee korkeimmalle. Siis suurin suhteellinen riskitiheys liittyy päivittäisissä toiminnoissa havaittuihin vaikeuksiin. Tämä suuri suhteellinen riskitiheys koskee kuitenkin vain pientä osaa havaintoyksiköistä, mahdollisesti vain yhtä. Noin 30-40 prosenttia havaintoyksiköistä on miehiä, sillä sukupuolen suhteellisen riskitiheyden kuvaajasta suunnilleen tuo osuus kuuluu miesten ryhmään. Riskitiheyssuhde miehille on 1.77. Kuvaajien mukaan painoindeksiin liittyvä suhteellinen riskitiheys on korkeintaan suunnilleen samaa luokkaa, ja nopeuteen sekä vaikeuksiin liittyvät suhteelliset riskitiheydet ylittävät kyseisen riskitiheyssuhteen noin 10 prosentilla havaintoyksiköistä. Kuolemisen riskitiheys miessukupuoleen liittyen näyttäisi olevan populaatiotasolla suurempi kuin esimerkiksi alipainoon liittyen.

Sukupuolen jakaumasta on huomattavissa myös se, että miehiä on ilmeisesti kuollut suhteessa enemmän jo ennen 75-vuoden ikää kuin naisia. Muuten sukupuolijakauman kuuluisi olla tasaisemmin jakautunut.

Luvussa 5.1 käytetään riskitiheyskuviota tutkittaessa, onko aineistossa olevilla poikkeavilla havainnoilla vaikutusta mallin tulkintaan. Suhteellisen riskitiheyden jakauman avulla löytää poikkeavat havainnot helposti, sillä kuvaajaan tulee pystysuoria kohtia etenkin häntiin.

Mikäli mallintajan mielestä esimerkiksi suhteellisen riskitiheyden jakauman kuvasta havaittu erittäin suuri suhteellinen riskitiheys vaikuttaa epäuskottavalta, mallia ei välttämättä kannata käyttää ainakaan niin suurilla tai pienillä kovariaatin arvoilla (Karvanen & Harrell, 2009). Kyseisestä kovariaatista voi myös kokeilla erilaista muunnosta, joita käsitellään luvussa 5.2. Mallin sopivuuden lisäksi kannattaa siis tarkastaa, millaisia tuloksia oletus suhteellisesta riskitiheydestä tuottaa mallin tulkinnan kannalta (Karvanen & Harrell, 2009). Riskitiheyskuvio on tähän oivallinen keino.

4.4 Luottamusvälit suhteelliselle riskitiheydelle

Suhteelliselle riskitiheydelle voidaan piste-estimaattien lisäksi piirtää luottamusvälit. Luottamusvälit risteävät vertailukohdassa ja ovat leveimmillään laitimmaisilla arvoilla. Mikäli estimoidun kertoimen luottamusväli ei sisällä nollaa, referenssi kuuluu vastaavalle suhteellisen riskitiheyden luottamusvälille vain vertailukohdassa. Päinvastaisessa tilanteessa referenssisuora jää luottamusvälien sisään. Luottamusvälejä piirrettäessä kannattaa selkeyden vuoksi esittää samaan kuvaan vain yhden kovariaatin suhteellisen riskitiheyden jakauma ja sen luottamusvälit. (Karvanen & Harrell, 2009.)

Collettin (2003) mukaan riskitiheyssuhteelle $\exp(\beta)$ voidaan laskea luottamusvälit käyttäen kertoimen $\hat{\beta}$ luottamusväliä. Riskitiheyssuhteen luottamusväli saadaan korvaamalla $\hat{\beta}$ luottamusvälinsä ala- ja ylärajalla riskitiheyssuhteen kaavassa. Tämä on jopa suositeltavampi tapa kuin riskitiheyssuhteen keskivirheen käyttäminen perinteisessä luottamusvälin kaavassa, sillä normaaliapproksimaatio toimii paremmin kertoimelle $\hat{\beta}$ kuin riskitiheyssuhteelle $\exp(\hat{\beta})$.

Suhteellinen riskitiheys $\exp\big(\beta(x-x_{\text{ref}})\big)$ lasketaan kiinnitetyillä kovariaatin arvolla x ja vertailukohdalla x_{ref}. Epävarmuus sisältyy siis kertoimen β estimaattiin. Suhteellisen riskitiheyden luottamusvälin laskemisessa voidaan käyttää estimoidun kertoimen $\hat{\beta}$ luottamusväliä samalla periaatteella kuin riskitiheyssuhteen kanssa. Vakiolla kertominen ei vaikuta normaaliapproksimaation toimimiseen. Luottamusvälin alaraja suhteelliselle riskitiheydelle on $\exp\big(\hat{\beta}_a(x-x_{\text{ref}})\big)$ ja yläraja $\exp\big(\hat{\beta}_y(x-x_{\text{ref}})\big)$, missä $(\hat{\beta}_a, \hat{\beta}_y)$ on luottamusväli kertoimelle β.

Koska suhteellinen riskitiheys faktorille pelkistyy riskitiheyssuhteeseen, kunkin tason suhteellisen riskitiheyden luottamusväli on kyseisen tason riskitiheyssuhteen luottamusväli.

Kuvaan 4 on piirretty monitasoisen faktorin sekä jatkuvan muuttujan luottamusvälit koodilla:

```
par(mar = c(2, 5, 4, 2) + 0.1); par(mfrow = c(2, 1))
rankhazardplot(malli1, data = gero, col = "darkgreen",
  ylim = c(0.7, 22),  yticks = c(0.7, 0.8, 0.9, 1, 2*(1:11)),
  yvalues = c(0.7, 0.8, 0.9, 1, 2*(1:4), 12, 20),
  refline = TRUE, ylab = "Suhteellinen riskitiheys",
  main= "Terveydentilan yhteys suhteelliseen riskitiheyteen",
  legendtext = "terv.tila", confint = TRUE)

rankhazardplot(malli2, data = gero, col = 3,
  yticks = c(0.2, 0.4, 0.6, 0.8, 1, seq(1.5, 4.5, by = .5)),
  yvalues = c(0.2, 0.6, 1, 2, 3, 4), refline = TRUE,
  main = "Nopeuden yhteys suhteelliseen riskitiheyteen",
  ylab = "Suhteellinen riskitiheys",
  legendtext = "nopeus", confint = TRUE, select = 3)
```

Ylemmässä kuvassa oleva terveydentila-faktori on peräisin mallista 1, jonka kertoimet on esitetty sivulla 19. Faktorin taso 1 on vertailutaso, joten sen luottamusväli surkastuu pisteeksi. Koska taso 2 ei eroa vertailutasosta, referenssi sisältyy tason 2 luottamusväliin. Tämän näkee helposti siitä, että mustalla katkoviivalla korostettu referenssi on tason 2 kohdalla luottamusvälin sisällä. Tasot 3 ja 4 eroavat vertailutasosta.

Alemman kuvan kävelynopeus-muuttuja on mallista 2. Kertoimet löytyvät sivun 22 taulukosta. Vertailukohta on kovariaatin mediaanissa, jossa siis luottamusvälin pituus on nolla. Luottamusväli on kapea lähellä vertailukohtaa, ja se levenee ääripäitä kohden. Koska muuttuja on tilastollisesti merkitsevä, luottamusväli sisältää arvon 1 ainoastaan vertailukohdassa.

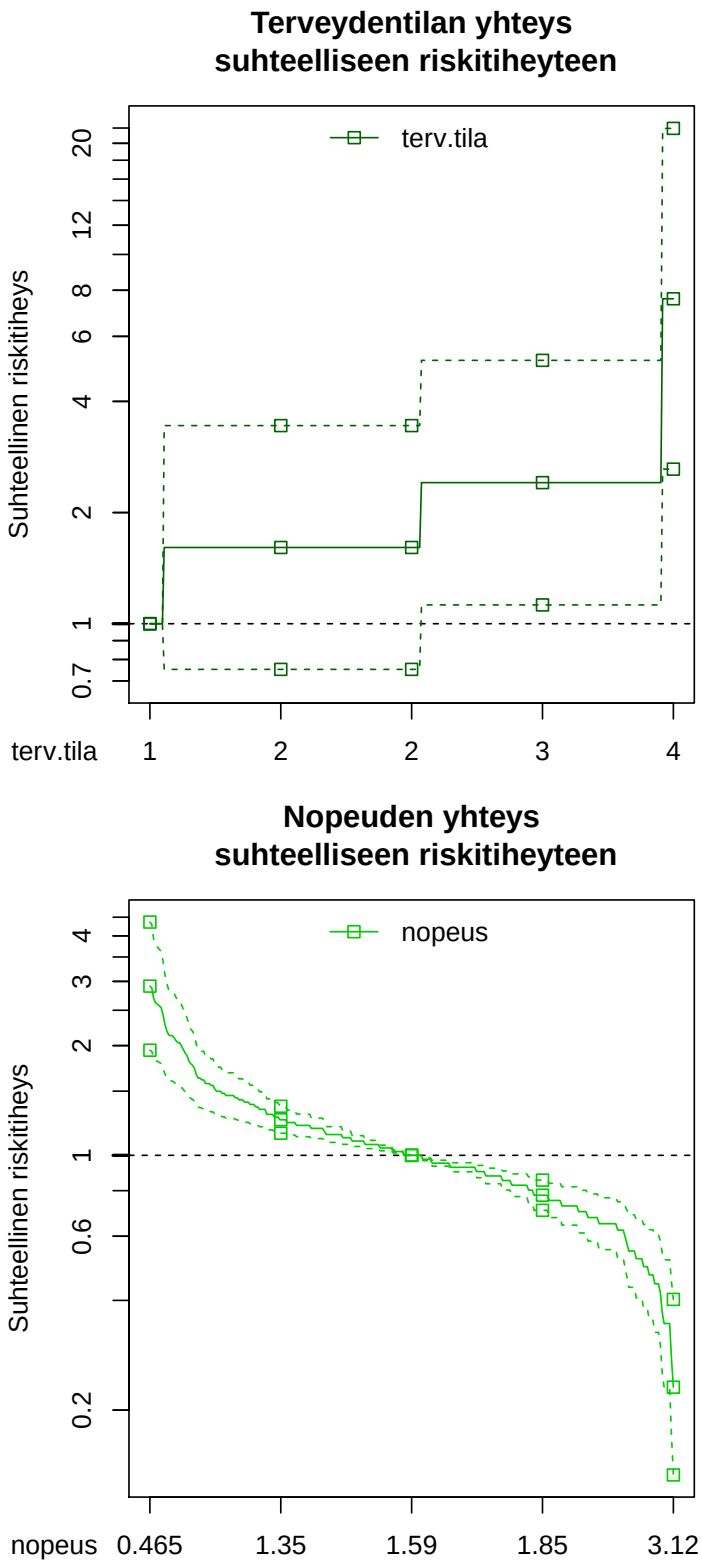
4.5 Ajassa muuttuvien kovariaattien malli

Tutkielmassa keskitytään esittelemään, kuinka riskitiheyskuviota voidaan käyttää Coxin suhteellisten riskitiheyksien mallin havainnollistamisessa. Riskitiheyskuviolla on mahdollista havainnollistaa muitakin malleja. Keskeistä on, että perusriskitiheys supistuu laskettaessa pois ja että kuvaa piirtäessä kukin havaintoyksikkö on edustettuna korkeintaan kerran.

Ensimmäisen ehdon voimassaolon kannalta on välttämätöntä, ettei perusriskitiheys riipu kovariaattien arvoista. Perusriskitiheyden supistuminen on tärkeää siksi, ettei sitä tarvitse estimoida. Toiseen ehtoon tulee kiinnittää erityistä huomiota, kun aineistossa on useita rivejä samasta havaintoyksiköstä. Tällöin on kyse start–stop-aineistosta, jossa kukin havaintoyksikkö on esitetty useana havaintona.

Jos potilas saa esimerkiksi siirrännäisen ajanhetkellä 21 ja sen jälkeen elää ajanhetkeen 345, on ensimmäinen aikaväli $(0, 21]$ ja toinen $(21, 345]$. Havaintoyksikkö sensuroituu ensimmäisen aikavälin lopussa ja kuolee toisen aikavälin lopussa. Syy kahtena rivinä esittämiseen on se, että alussa potilaalla ei ole siirrännäistä mutta toisen välin ajan on. Tämä muutos saadaan huomioitua start–stop-aineiston avulla. (Therneau & Grambsch, 2000.)

Ikivihreät-aineistossa on muuttujia, joiden arvot on mitattu tutkimuksen alussa ja uudelleen 80 vuoden iässä. Kovariaatin arvosta saadaan siis päivitetty tieto viisi vuotta tutkimuksen alkamisesta. Tällaisessa tilanteessa voidaan käyttää ajassa muuttuvien kovariaattien mallia, joka on eräs Coxin malli. Tässä mallissa suhteellisen riskitiheyden oletus pätee vain jokaisella ajanhetkellä t, sillä kovariaattien arvojen muuttuessa suhteellinen riskitiheys kahden yksilön välillä voi muuttua, eikä se siis pysy samana koko aikaa. (Collett, 2003.)



Kuva 4: Ylhäällä mallin 1 terveydentila-faktorin luottamusvälit. Vertailutasona epätavallisen hyvä terveydentila. Mitä suuremman arvon faktori saa, sitä huonompi terveydentila henkilöllä on. Alhaalla luottamusvälit jatkuvalle kävelynopeus-muuttujalle mallista 2. Vertailukohtana kovariaatin mediaani.

Taulukko 5: Coxin mallin kertoimet, riskitiheyssuhteet ja merkitsevyyden testaus Ikivihreät-aineistoon sovitetulle mallille, jossa ajassa muuttuvia kovariaatteja. Mallin n = 459, havainnoista 283 kuuluvat aikavälille ennen kovariaattien päivittämistä ja 176 kovariaattien päivittämisen jälkeen.

Muuttuja	$\hat{\beta}$	$\exp(\hat{\beta})$	$\text{se}(\hat{\beta})$	p
sukupuoli	0.727	2.07	0.155	<0.001
painoindeksi	-0.069	0.93	0.019	<0.001
nopeus	-1.101	0.33	0.203	<0.001
vaikeudet	0.046	1.05	0.014	<0.001

Kun elinaika on mitattu tutkimuksen alusta, riskitiheysfunktio henkilölle i hetkellä t on

$$h_i(t) = h_0(t)\exp(\beta_1 x_{i1}(t) + \ldots + \beta_p x_{ip}(t)) = h_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{x}_i(t)),$$

missä $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ ja $\boldsymbol{x}_i(t)$ on henkilön i kovariaattien arvot vektorina hetkellä t. Perusriskitiheys kuvaa riskitiheyttä henkilölle, jolla kaikki kovariaatin arvot ovat nollia alkuhetkellä, eivätkä muutu seurannan aikana. (Collett, 2003.)

Nyt suhteellisen riskitiheyden arvo muuttuu ajassa, ja siksi sen jakauma voidaan kuvata vain tietyillä ajanhetkillä. Vertailukohtana voidaan käyttää mitä tahansa arvoa, mutta sen tulkitaan olevan samalla ajanhetkellä t mitattu arvo kuin piirrettävät arvot ovat. Tällöin perusriskitiheys on sama vertailukohdalle ja muille arvoille, ja se supistuu pois. Kuvaa piirrettäessä valitaan tietyt kovariaatin arvot, esimerkiksi tutkimuksen alussa mitatut. Mikäli kuvien perusteella halutaan vertailla suhteellisen riskitiheyden jakaumien muutosta mittausten välillä, kannattaa käyttää samaa vertailukohtaa kaikissa kuvissa, jotta lasketut suhteelliset riskitiheydet ovat vertailukelpoisia, eli samalla kovariaatin arvolla saadaan sama suhteellinen riskitiheys joka kuvaan.

Ikivihreät-aineistoon on sovitettu ajassa muuttuvien kovariaattien malli, jossa selittävinä muuttujina on käytetty sukupuolta, painoindeksiä, päivittäisissä toiminnoissa koettujen vaikeuksien indeksiä sekä kävelynopeutta. Aineiston luominen on esitetty liitteessä E. Sovitetun mallin kertoimet ja tilastollinen testaus on esitetty taulukossa 5. Malli on sovitettu seuraavalla koodilla:

```
aikamalli <- coxph(Surv(time1, time2, status) ~ sukup + bmi +
    nopeus + vaik, data = gero_aika, x = TRUE)
```

Kuvissa 5 ja 6 on piirretty taulukossa 5 esitetyn mallin mukaisia riskitiheyskuvioita eri aineistoilla. Kaikissa on käytetty vertailukohtina tutkimuksen alussa mitattuja kovariaattien mediaaneja. Kuvassa 5 ylhäällä on aineistona tutkimuksen alussa mitatut kovariaattien arvot kaikista havaintoyksiköistä. Nähdään, että suurimmat suhteelliset riskitiheydet liittyvät nopeuteen ja päivittäisissä toiminnoissa ilmeneviin vaikeuksiin siten, että hitaimmilla ja eniten

vaikeuksista kärsivillä on suurimmat riskitiheydet verrattuna mediaaneihin. Kuitenkin suuriin suhteellisiin riskitiheyksiin liittyviä kovariaatin arvoja ilmenee aineistossa melko vähän. Vain noin kymmenen prosenttia molempien kovariaattien arvoista ylittää sukupuoleen liittyvän riskitiheyssuhteen. Miehillä on yli kaksinkertainen riskitiheys naisiin verrattuna. Kuvaajasta nähdään myös, että 75-vuotiaista henkilöistä noin 30–40 prosenttia on miehiä.

Tätä kuvaa voidaan verrata sekä kuvassa 5 alhaalla olevaan kuvaan, jossa aineistona on 80 vuoden iässä mitatut kovariaattien arvot että kuvassa 6 ylhäällä olevaan kuvaan, johon on valittu ennen 80 vuoden mittauksia kuolleet henkilöt.

Vertailussa ensin mainittuun huomataan, että sukupuolen, painoindeksin ja vaikeuksien jakaumat näyttävät likimain samoilta kuvien välillä. Ainoa havaittava muutos on kävelynopeudessa. Arvo 1.59 m/s, joka on ensimmäisen mittauskerran mediaani, on likimain toisen mittauskerran yläkvartiili (1.61 m/s). Koehenkilöt siis hidastuvat mittauskertojen välillä. Toinen mahdollisuus olisi se, että nopeammat henkilöt kuolisivat ennen 80 vuoden mittauksia, mutta näin asia ei ole: se nähdään yläkuvasta 6. Taas näyttää siltä, että kuolleiden jakaumat muistuttavat koko aineiston jakaumia muuten, mutta nopeuden jakauma on erilainen: hitaat kävelijät näyttävät olevan yliedustettuna kuolleiden joukossa. Tämä antaa viitettä siitä, että 75–79 vuoden iässä kävelynopeus selittäisi näistä muuttujista vahvimmin kuolleisuutta.

Kuvassa 6 alhaalla on piirretty jakaumat iässä 80–84 kuolleiden toisen mittauksen kovariaattien arvoista. Tässä iässä miehiä kuolee noin 60 prosenttia, mikä on melkein tuplasti miehien osuus aineistosta. Myös nopeuteen liittyvä suhteellisen riskitiheyden jakauma korostuu entisestään.

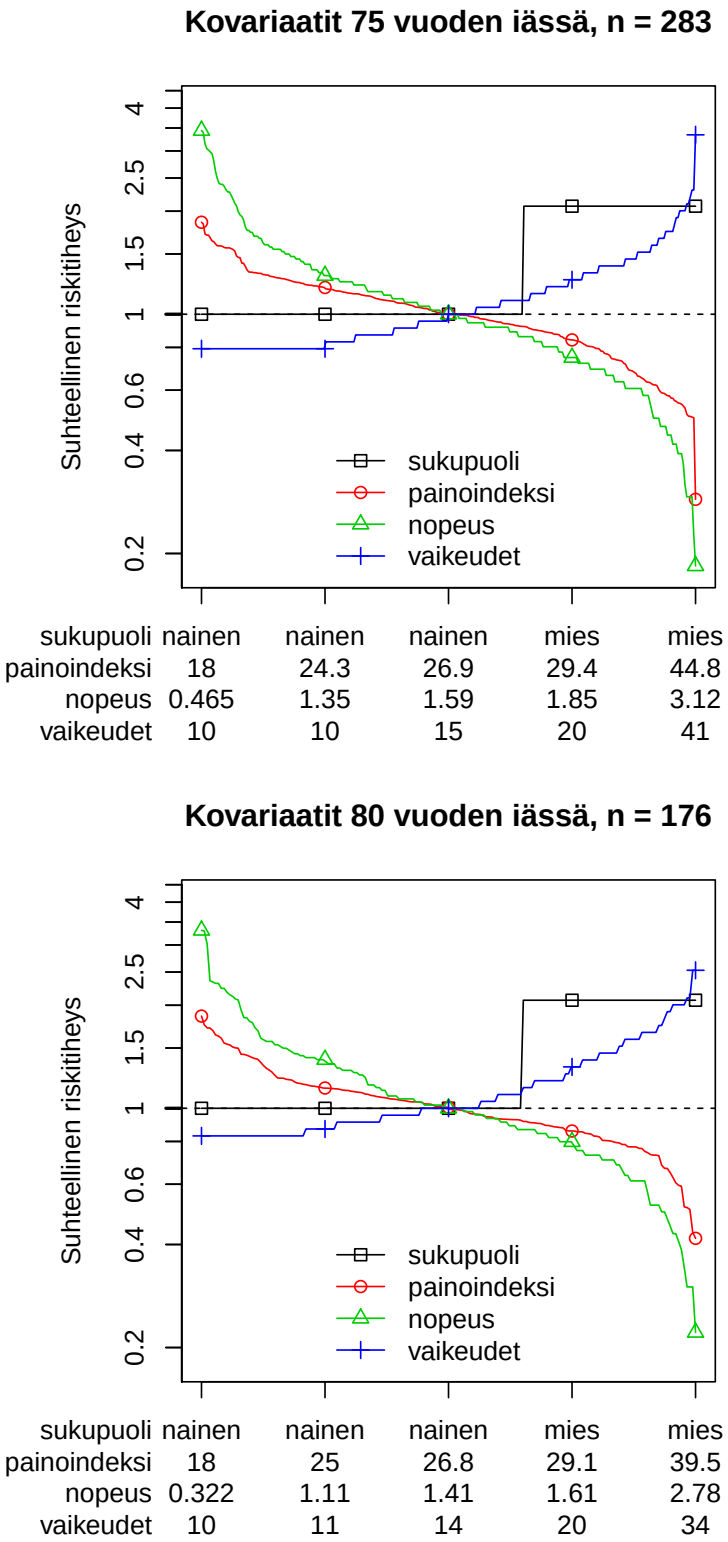
Kuvassa 5 ylhäällä olevan kuvan suhteelliset riskitiheydet tulkitaan siten, että esimerkiksi hitaimman henkilön kuolemisen riskitiheys verrattuna mediaanin riskitiheyteen on noin 3.5-kertainen tutkimuksen ensimmäisen viiden vuoden aikana. Alhaalla olevan kuvan mukaan hitaimman henkilön kuolemisen riskitiheys on tutkimuksen loppuajan noin 4-kertainen vertailukohdan riskitiheyteen ajatellen, että vertailukohta ei ole muuttunut ensimmäisten viiden vuoden jälkeen. Kuitenkaan kyseessä ei ole riskitiheys verrattuna tutkimuksen alussa mitattuun vertailukohdan riskitiheyteen, sillä perusriskitiheys muuttunee tutkimuksen kuluessa.

Kuvan 5 piirtokoodi

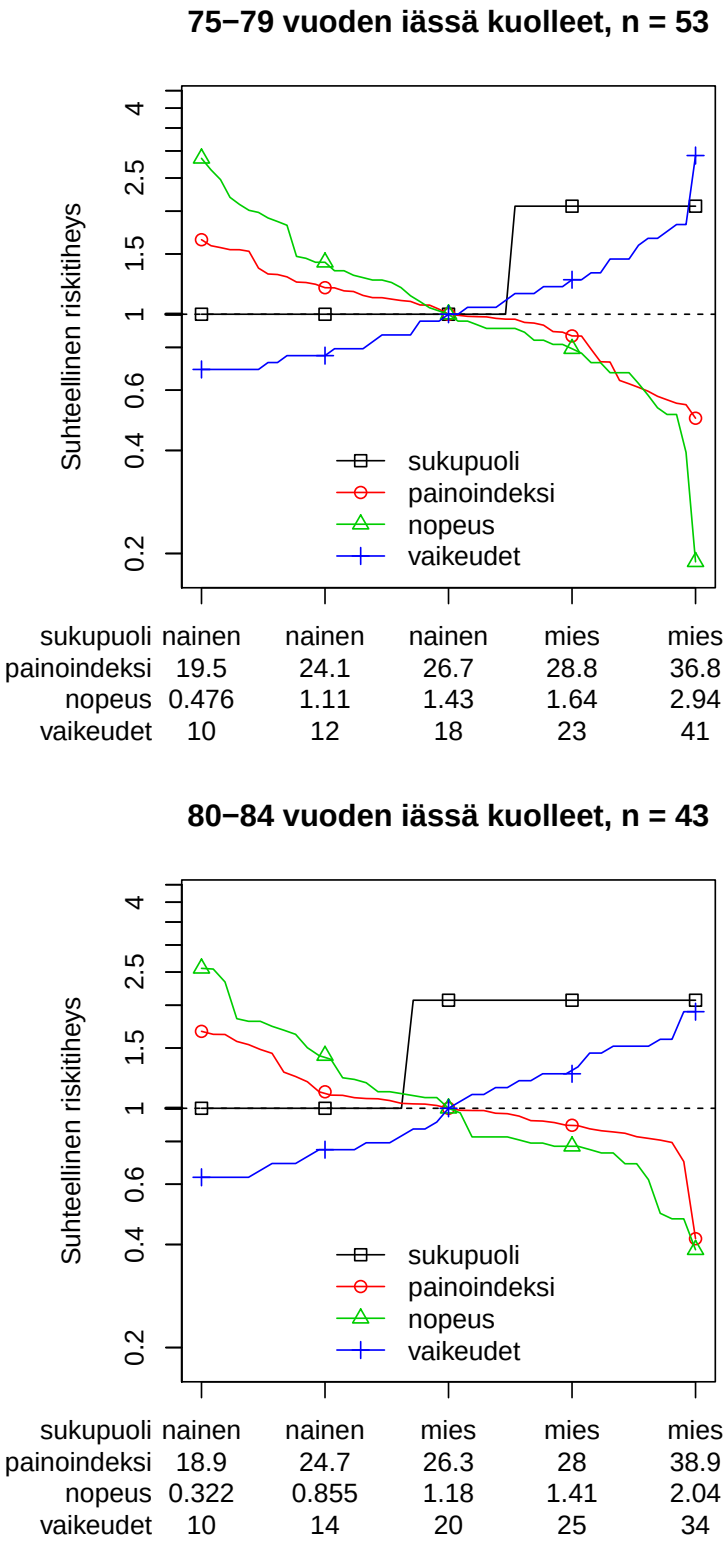
```
par(mar = c(5, 6, 4, 2) + 0.1); par(mfrow = c(2, 1))
rankhazardplot(aikamalli, data = gero_aika[gero_aika$time1 == 0,],
  ylim = c(0.18, 4.1),  ylab = "Suhteellinen riskitiheys",
  legendlocation = "bottom", refline = TRUE,
  main = "Kovariaatit 75 vuoden iässä, n = 283",
  legendtext = c("sukupuoli", "painoindeksi", "nopeus", "vaikeudet"))

rankhazardplot(aikamalli, legendlocation = "bottom",
  data = gero_aika[gero_aika$time1 == 1826,], refline = TRUE,
  ylim = c(0.18, 4.1), ylab = "Suhteellinen riskitiheys",
  main = "Kovariaatit 80 vuoden iässä, n = 176",
  legendtext = c("sukupuoli", "painoindeksi", "nopeus", "vaikeudet"))
```

Kuvan 6 piirtokoodi

```
par(mar = c(5, 6, 4, 2) + 0.1); par(mfrow = c(2, 1))
rankhazardplot(aikamalli, legendlocation = "bottom",
  data = gero_aika[gero_aika$time2 < 1826,],
  ylim = c(0.18, 4.1), refline = TRUE,
  main = "75-79 vuoden iässä kuolleet, n = 53",
  ylab = "Suhteellinen riskitiheys",
  legendtext = c("sukupuoli", "painoindeksi", "nopeus", "vaikeudet"))

rankhazardplot(aikamalli, legendlocation = "bottom",
  data = gero_aika[gero_aika$time2 > 1826 & gero_aika$time2 < 3652,],
  ylim = c(0.18, 4.1), ylab = "Suhteellinen riskitiheys",
  main = "80-84 vuoden iässä kuolleet, n = 43", refline = TRUE,
  legendtext = c("sukupuoli", "painoindeksi", "nopeus", "vaikeudet"))
```



Kuva 5: Ajassa muuttuvien kovariaattien Coxin mallin mukaiset riskitiheyskuviot kovariaattien ensimmäisellä mittauskerralla (75-vuotiaat) ja toisella mittauskerralla (80-vuotiaat). Vertailukohtana ovat kovariaattien mediaanit 75 vuoden iässä.



Kuva 6: Ajassa muuttuvien kovariaattien Coxin mallin mukaiset riskitiheyskuviot iässä 75–79 ja 80–85 kuolleille. Vertailukohtana ovat kovariaattien mediaanit 75 vuoden iässä.

5 Coxin mallin diagnostiikkaa riskitiheyskuviota käyttäen

Riskitiheyskuvio näyttää mallin antamat ennusteet suhteelliselle riskitiheydelle kovariaattien eri arvoilla. Kuvio ei kuitenkaan vertaa ennusteita havaittuihin arvoihin, joten mallin hyvyyttä ei voida tarkastella tästä näkökulmasta. Tähän sopivia keinoja on esitelty luvussa 2.4. Diagnostiikkaan kuuluu kuitenkin muutakin kuin mallin sopivuus aineistoon, kuten vaikuttavien havaintojen tutkiminen sekä erilaisten muunnosten tarkastelu kovariaateille. Näihin riskitiheyskuvio tuo erilaisen näkökulman kuin muut olemassa olevat diagnostiikkakeinot.

5.1 Poikkeavien havaintojen vaikutus

Vaikuttava havainto tarkoittaa, että yhdellä havaintoyksiköllä on suuri merkitys siihen, millaiset kertoimet malli antaa. Siis jos vaikuttava havainto poistetaan aineistosta, mallin antamat tulokset muuttuvat merkittävästi. Havaintojen vaikuttavuutta voi tutkia delta-betojen avulla (Cain & Lange, 1984). Nimessä delta kuvaa muutosta ja beta estimoitavaa kerrointa. Jokaisen havaintoyksikön jokaiselle kovariaatille lasketaan arvo, joka approksimoi sitä, kuinka paljon kovariaatin kerroin muuttuu, jos havainto poistetaan aineistosta. Negatiiviset arvot kuvaavat kertoimen kasvua ja positiiviset kertoimen pienenemistä. (Collett, 2003.)

Delta-betojen laskemisen jälkeen tulee tarkastella, onko jonkin havaintoyksikön arvo muita suurempi itseisarvoltaan. Tämän voi tehdä esimerkiksi piirtämällä kovariaateittain kuvan, jossa vaaka-akselilla on havaintoyksikön rivinumero ja pystyakselilla delta-beta-arvo. Mikäli jokin havainto poikkeaa muista paljon delta-betojen perusteella, voidaan sen poistamista aineistosta harkita. (Collett, 2003.)

Poikkeavat havainnot voivat olla tällaisia vaikuttavia havaintoja. Poikkeavalla havainnolla tarkoitetaan sitä, että havaintoyksiköllä on jonkin kovariaatin arvo erityisen suuri tai pieni verrattuna muihin mallissa oleviin havaintoyksikköihin. Mikäli poikkeava havainto muuttaa mallin kertoimia huomattavasti, voi sen poistaminen mallista olla aiheellista. Kertomien muutosta voi tarkastella sovittamalla mallin aineistoon, josta on poistettu poikkeavan havainnon sisältävä havaintoyksikkö. Sitten tämän mallin kertoimia verrataan aikaisempaan malliin.

Poikkeavat havainnot voidaan nähdä riskitiheyskuviosta. Mikäli kovariaatin suhteellisen riskitiheyden jakauma on hännästä jyrkästi laskeva tai nouseva, viittaa se poikkeavan havainnon ilmenemiseen. Esimerkiksi kuvassa 3 (sivu 23) nähdään lähes pystysuora kuvaaja sekä painoindeksin että vaikeudetmuuttujan maksimeissa. Painoindeksin maksimi on 44.8, ja tarkemman tarkastelun perusteella toiseksi suurin arvo on 36.9.

Sovitetaan Ikivihreät-aineistoon taulukossa 4 esitettyä mallia vastaava mal-

Taulukko 6: Coxin mallin kertoimet, riskitiheyssuhteet ja merkitsevyyden testaus Ikivihreät-aineistoon sovitetulle mallille numero 3, josta on poistettu painoindeksin mukaan poikkeava havainto. Otoskoko on 282.

Muuttuja	$\hat{\beta}$	$\exp(\hat{\beta})$	$\text{se}(\hat{\beta})$	p
sukupuoli_mies	0.570	1.77	0.139	<0.001
painoindeksi	-0.073	0.93	0.018	<0.001
nopeus	-0.950	0.39	0.184	<0.001
vaikeudet	0.053	1.05	0.012	<0.001

li ilman havaintoyksikköä, jolla on havaittu muista poikkeava painoindeksin arvo. Aineiston muokkaus on tehty koodilla:

```
gero2 <- gero[gero$bmi75 < 44, ]
aika2 <- Surv(gero2$eloaika, gero2$delta)
malli3<- coxph(aika2 ~ sp + bmi75 + nopeus75 + vaik75,
  data = gero2, x = TRUE)
```

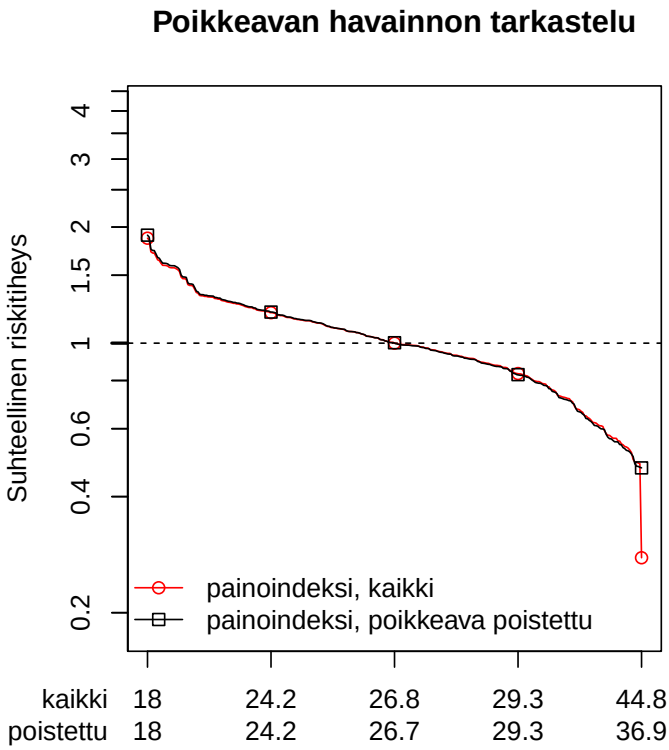
Tämän mallin tulokset on esitetty taulukossa 6. Huomataan, että kertoimissa ei ole suurta eroa verrattuna malliin 2, joten poikkeava havainto ei ole vaikuttava. Piirretään kuitenkin vertailua varten painoindeksin suhteellisen riskitiheyden jakaumat samaan kuvaan molemmista malleista:

```
bmi_coefs <- c(malli2$coef["bmi75"], malli3$coef["bmi75"])
bmi_x <- data.frame(gero$bmi75, c(gero2$bmi75, NA))

par(mar = c(3, 6, 4, 2) + 0.1))
rankhazardplot(x = bmi_x, coefs = bmi_coefs,  na.rm = FALSE,
  col = 2:1, pch = 1:0, ylim = c(0.18, 4.1),
  legendtext = c("painoindeksi, kaikki",
  "painoindeksi, poikkeava poistettu"),
  axistext = c("kaikki", "poistettu"), axistextpos = -0.08,
  main = "Poikkeavan havainnon tarkastelu",
  legendlocation = "bottomleft", refline = TRUE ,
  ylab = "Suhteellinen riskitiheys")
```

Kuvasta 7 nähdään, että ainoa silmällä nähtävä ero liittyy poikkeavan havainnon puuttumiseen toisen jakauman kuvaajasta.

5.2 Kovariaattien muodon tutkiminen

Kovariaatin muunnos voi parantaa mallin sopivuutta aineistoon (Collett, 2003). Luvussa 2.4 mainittuja martingaaliresiduaaleja käyttäen voidaan tutkia, mikä muunnos sopisi mallissa käytettävälle kovariaatille parhaiten. Tätä varten



Kuva 7: Riskitiheyskuvio, jossa on esitetty vain painoindeksiin liittyvät suhteelliset riskitiheydet Coxin suhteellisten riskitiheyksien malleista, joissa on mukana myös sukupuoli, nopeus ja päivittäisissä toimissa koettujen vaikeuksien pistemäärä. Mallit poikkeavat toisistaan siten, että mustalla piirretty kuvaaja on tehty mallista, josta on poistettu painoindeksiltään poikkeava havaintoyksikkö. Vertailukohtina on käytetty molempien muuttujien mediaaneja.

martingaaliresiduaalit lasketaan mallista, jossa ei ole selittäviä muuttujia. Nämä martingaaliresiduaalit piirretään kovariaattien arvojen funktiona, ja pistejoukko näyttää tarvittavan muunnoksen muodon. (Therneau ym., 1990.)

Muunnoksen käyttö kovariaatissa voi vaikuttaa myös kovariaatin tulkintaan. Muunnosten vaikutusta riskitiheyteen voidaan vertailla sovittamalla useita malleja ja tarkastelemalla niiden riskitiheyskuvioita. Karvanen ja Harrell (2009) painottavat, että tällöin muiden kuin muunnetun kovariaatin tulee olla samoja jokaisessa mallissa, sillä estimoitu riskitiheyssuhde riippuu sekä muunnoksesta että muista mallissa olevista kovariaateista. Riskitiheyskuvioon piirretään ainoastaan eri muunnosten kuvaajat vertailua varten.

Mallin 2 riskitiheyskuviosta kuvassa 3 huomataan, että vaikeudet-muuttujan jakauma on oikealle vino: ensimmäiset 50 prosenttia havainnoista ovat välillä 10–15, jälkimmäiset välillä 15–41. Maksimissa havaitaan myös poikkeava havainto. Logaritmimuunnos lyhentää oikeaa häntää ja voisi parantaa mallin sopivuutta aineistoon. Tutkitaan seuraavaksi, miten logaritmimuunnos vaikuttaa

Taulukko 7: Coxin mallin kertoimet, riskitiheyssuhteet ja merkitsevyyden testaus Ikivihreät-aineistoon sovitetulle mallille numero 4, jossa on käytetty vaikeudet-muuttujalle e-kantaista logaritmimuunnosta. Otoskoko on 283.

Muuttuja	$\hat{\beta}$	$\exp(\hat{\beta})$	$\text{se}(\hat{\beta})$	p
sukupuoli_mies	0.584	1.79	0.139	<0.001
painoindeksi	-0.069	0.93	0.017	<0.001
nopeus	-0.934	0.39	0.185	<0.001
log(vaikeudet)	0.922	2.51	0.216	<0.001

vaikeudet-muuttujan suhteellisen riskitiheyden jakaumaan.

Sovitetaan vertailtava malli, jossa on muuten samat muuttujat kuin mallissa 2. Ainoastaan vaikeudet-muuttujalle on käytetty e-kantaista logaritmimuunnosta:

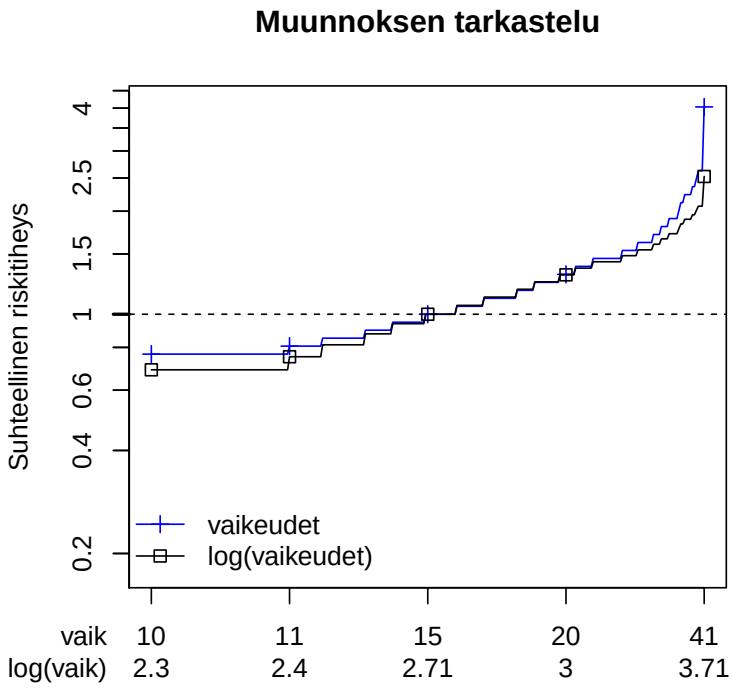
```
aika <- Surv(gero$eloaika, gero$delta)
malli4 <- coxph(aika ~ sp + bmi75 + nopeus75 + log(vaik75),
  data = gero, x = TRUE)
```

Mallin tulokset on esitetty taulukossa 7. Vertailuna mallin 2 kertoimiin sivulla 22 huomataan, että vaikeudet muuttujan kerroin on muuttunut mallien välillä arvosta 0.054 arvoon 0.922, eli 17-kertaiseksi. Muissa kertoimissa ei ole suuria muutoksia. Merkittävä muutos kertoimessa ei tarkoita automaattisesti muutosta suhteellisen riskitiheyden jakaumassa: Suhteellisen riskitiheyden jakaumaa piirrettäessä logaritmimuunnetulle kovariaatille käytetään logaritmimuunnettuja arvoja, jotka ovat pienempiä kuin kovariaatin alkuperäiset arvot. Lisäksi referenssiarvo on eri.

Piirretään jakaumat samaan kuvaan. Käytetään malleista poimittuja kertoimia piirtämiseen. Datana annetaan sekä alkuperäiset että muunnetut kovariaatin arvot.

```
vaik_coefs <- c(malli2$coef["vaik75"], malli4$coef["log(vaik75)"])
vaik_x <- data.frame(gero$vaik75, log(gero$vaik75))

par(mar = c(3, 5, 4, 2) + 0.1))
rankhazardplot(x = vaik_x, coefs = vaik_coefs,  na.rm = FALSE,
  col = c(4, 1), pch = c(3, 0), ylim = c(0.18, 4.1),
  legendtext = c("vaikeudet", "log(vaikeudet)"),
  axistext = c("vaik", "log(vaik)"),   axistextpos = -0.08,
  legendlocation = "bottomleft", main = "Muunnoksen tarkastelu",
  refline = TRUE, ylab = "Suhteellinen riskitiheys")
```

Kuvassa 8 on esitetty kuvasta 3 vaikeudet-muuttujan ja sen logaritmimuunnos mallista, jossa on mukana myös sukupuoli ja nopeus. Molempien muuttujien suhteellisten riskitiheyksien jakaumat ovat melko samanlaiset. Minimin



Kuva 8: Riskitiheyskuvio, jossa on esitetty vain vaikeudet-muuttujan ja sen logaritmimuunnoksen antamat suhteelliset riskitiheydet Coxin malleista, joissa on mukana myös sukupuoli, painoindeksi ja nopeus. Vertailukohtana on käytetty kovariaattien mediaaneja.

ja alakvartiilin välillä logaritmimuunnoksen suhteellinen riskitiheys on hieman pienempi kuin alkuperäisen muuttujan. Maksimissa suhteellinen riskitiheys ei kasva yhtä suureksi kuin alkuperäisellä muuttujalla. Koska suhteellisen riskitiheyden jakaumissa ei ole tulkinnallisesti erityisen paljon eroa, voidaan muunnoksen valinnassa kääntyä kokonaan martingaaliresiduaalien puoleen.

Tutkitaan seuraavaksi, onko riittävää laittaa painoindeksi malliin lineaarisena, vai olisiko muuttujalle tarpeellista käyttää monimutkaisempaa muotoa. On esimerkiksi mahdollista, että sekä alhainen että korkea painoindeksi lisää kuolemisen riskiä. Tätä varten sovitetaan taas mallia 2 vastaava malli siten, että painoindeksiin sovitetaan P-spline (Eilers & Marx, 1996). Spline on lokaalisti määritettävä funktio, jonka avulla voi tutkia kovariaatin muotoa. Malli on sovitettu koodilla:

```
aika <- Surv(gero$eloaika, gero$delta)
malli5 <- coxph(aika ~ sp + pspline(bmi75) + nopeus75 + vaik75,
  data = gero, x = TRUE)
```

Taulukko 8: Coxin mallin kertoimet, riskitiheyssuhteet ja merkitsevyyden testaus Ikivihreät-aineistoon sovitetulle mallille numero 5, jossa painoindeksille on tehty P-spline-sovitus. Otoskoko on 283.

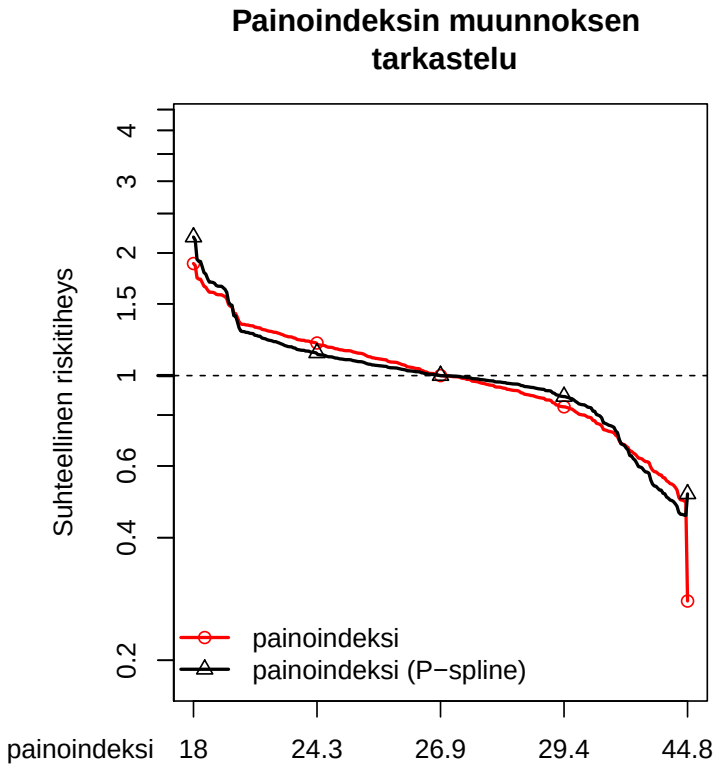
Muuttuja	$\hat{\beta}$	$\exp(\hat{\beta})$	$\text{se}(\hat{\beta})$	p
sukupuoli_mies	0.564	1.76	0.140	<0.001
painoindeksi, lineaarinen	-0.071	0.93	0.017	<0.001
painoindeksi, epälineaarinen				0.7
nopeus	-0.938	0.39	0.185	<0.001
vaikeudet	0.055	1.06	0.013	<0.001

Taulukosta 8 nähdään, että painoindeksin epälineaarinen osa ei ole tilastollisesti merkitsevä. Tulosten mukaan on siis riittävää pitää painoindeksi lineaarisena mallissa. Verrataan kuitenkin suhteellisen riskitiheyksien jakaumia lineaariselle ja P-spline-painoindeksille:

```
tulos2 <- rankhazardplot(malli2, data = gero, return = TRUE,
  draw = FALSE)
tulos5 <- rankhazardplot(malli5, data = gero, return = TRUE,
  draw = FALSE)
bmi_x <- data.frame(tulos2$x[,"bmi75"],  tulos5$x[,"bmi75"])
bmi_xp <- data.frame(tulos2$xp[,"bmi75"],
  tulos5$xp[,"pspline(bmi75)"])
bmi_ref <- c(tulos2$ref["bmi75"], tulos5$ref["pspline(bmi75)"])

par(mar = c(2, 6, 4, 2) + 0.1)
rankhazardplot(x = bmi_x, xp = bmi_xp,  refvalues = bmi_ref,
  col = c(2, 1, 4), pch = c(1, 2, 3), ylim = c(0.18, 4.1),
  legendtext = c("painoindeksi","painoindeksi (P-spline)"),
  axistext = c("painoindeksi","painoindeksi"),
  main = "Painoindeksin muunnoksen tarkastelu",
  legendlocation = "bottomleft", na.rm = FALSE,
  ylab = "Suhteellinen riskitiheys",
  lwd = 2, axistextpos = -0.08, refline = TRUE)
```

Kuvasta 9 nähdään, että jakaumat eroavat käyttäytymiseltään vain poikkeavan havainnon kohdalla: spline-sovitus antaa poikkeavalle havainnolle hieman suuremman suhteellisen riskitiheyden kuin sitä edeltävälle havainnolle. Riskitiheyskuvio tukee siis tulosta lineaarisen muodon riittävyydestä painoindeksille.

Kolmas esimerkki muunnosten tarkastelusta löytyy liitteestä C.7. Kuva 17 sivulla 88 esittää bilirubiinimuuttujasta alkuperäisen, logaritmimuunnetun sekä sovitetun P-splinen suhteellisen riskitiheyden jakaumat. Jokaisessa käytetyssä mallissa on muuten samat kovariaatit, vain bilirubiinin muunnos vaihtelee.



Kuva 9: Riskitiheyskuvio, jossa on esitetty vain lineaarisen ja P-spline sovitetun painoindeksin suhteellisen riskitiheyden jakaumat. Malleissa on ollut mukana selittävinä muuttujina myös sukupuoli, nopeus ja päivittäisissä toiminnoissa koetut vaikeudet. Vertailukohtana on käytetty painoindeksin mediaania.

Kuvasta nähdään, että logaritmi- ja spline-muunnos tuottavat samankaltaiset suhteellisen riskitiheyden jakaumat. Alkuperäisen bilirubiini-kovariaatin suhteellisen riskitiheyden jakauma taas poikkeaa tulkinnaltaan niistä. Sen mukaan kuolemisen riski on lähes vakio minimistä yläkvartiiliin, ja vasta erittäin suurilla bilirubiiniarvoilla kuoleman riski kasvaa yli kaksinkertaiseksi mediaaniin nähden. Logaritmi- ja spline-muunnos antavat enemmän lineaarista muistuttavan jakauman. Mikä muunnoksista tuottaa tulkinnallisesti parhaan jakauman, on enemmän sisällöllinen kuin tilastotieteellinen kysymys.

6 Rankhazard-paketin toiminta

Tässä luvussa esitellään `rankhazard`-paketin (Karvanen & Koski, 2014) rakenne. Paketin toiminta kiteytyy `rankhazardplot`-funktion käyttämiseen, ja luvussa 6.2 käydään läpi eri tavat piirtää riskitiheyskuvio sekä luottamusvälit suhteellisen riskitiheyden jakaumalle. Tätä ennen luvussa 6.1 syvennytään `rankhazardplot`-funktion toimintaperiaatteeseen teknisestä näkökulmasta. Lopuksi luvussa 6.3 esitellään versioiden 0.8-1 ja 1.0 välillä toteutetut päivitykset.

Paketti `rankhazard` (Karvanen & Koski, 2014) sisältää kuusi funktiota, mutta käyttäjän tarvitsee kutsua vain `rankhazardplot`-funktiota. Funktiot on kirjoitettu siten, että niiden käyttäminen olisi tyyliltään samanlaista kuin muiden R-ympäristön funktioiden käyttö: yksinkertaista mutta monipuolista. Yksinkertaisuus tarkoittaa, että kuvan saa piirrettyä vain muutamia argumentteja käyttäen. Monipuolisuus taas tarkoittaa sitä, että käyttäjällä on mahdollisuus muuttaa lähes kaikkea tarpeellista kuvassa: esimerkiksi tekstejä, värejä, tyylejä ja asteikoita. Lisäksi argumenttien nimeämisessä on käytetty muista funktioista tuttuja argumenttien nimiä ja nimeämistyylejä.

Olen kirjoittanut dokumentaation `rankhazardplot`-funktiolle. Se on julkaistu Internetiin osoitteeseen `http://cran.r-project.org/web/packages/rankhazard/rankhazard.pdf`. Dokumentaatio on tutkielman liitteessä B. Dokumentaatiossa esitellään funktion ja sen argumenttien käyttötarkoitusta. Liitteessä A syvennetään dokumentaatiota käsittelemällä `rankhazardplot`-funktion toimintaa. Lisäksi käydään läpi kaikkien argumenttien käyttötarkoitus ja luodaan katsaus siihen, kuinka `rankhazardplot`-funktion tuottamaa kuvaa voidaan muokata `par`-funktiota käyttäen. Koodin kautta perehdytään kuvan piirtämiseen ja siinä tarvittavien arvojen laskemiseen. Liitteessä esitellään myös funktion palauttamat arvot ja tulosteet. Paketin `rankhazard` sisältämien funktioiden koodit versiosta 1.0 löytyvät liitteestä D.

Dokumentaation esimerkeistä on tehty oma liitteensä C. Esimerkkeihin on lisätty mukaan tulosteet ja kuvat. Näitä esimerkkejä käytetään tutkielmassa kuvaamaan funktion käyttömahdollisuuksia ja -tapoja. Kuvan 22 esimerkkiä on muokattu hieman julkaistusta versiosta: malliin on lisätty muuttuja `cluster(id)`, joka ottaa keskivirheiden laskemisessa huomioon sen, että havaintoyksiköistä voi olla useita rivejä aineistossa (Therneau & Grambsch, 2000). Klusterin lisääminen vaikuttaa vain keskivirheisiin, joten kuva ei muutu, mutta esimerkki on monipuolisempi.

Dokumentaation esimerkissä kuvassa 17 on käytetty spline-muunnosta. Therneau ja Grambsch (2000) ovat esitelleet useita erilaisia tapoja estimoida splinea: esimerkiksi regressio- ja kuutio-splineja sekä luonnollisia, silottavia ja penalisoituja splineja. Esimerkissä on käytetty P-splinea (Eilers & Marx, 1996).

P-splinea on käytetty koodin testaamiseen esimerkkinä monimutkaisesta muunnoksesta, jolle voi antaa useita argumentteja, esimerkiksi vapausasteet. Mitä enemmän vapausasteita on, sitä monimutkaisempia käyriä voidaan sovit-

taa (Therneau & Grambsch, 2000.). On ollut tärkeää testata funktiota tällaisella muuttujalla, jotta varmistetaan `rankhazardplot`-funktion toiminta erikoisissakin tapauksissa. Näille monimutkaisille muunnoksille ei kuitenkaan voida tällä hetkellä laskea luottamusvälejä.

6.1 Funktion `rankhazardplot` toimintaperiaate

Paketti `rankhazard` sisältää yhden kutsuttavan funktion, `rankhazardplot`, jolla voi piirtää suhteellisen riskitiheyden jakaumia. Funktio `rankhazardplot` käyttää S3-järjestelmää, jossa on oma funktionsa luokille `coxph` ja `cph`. S3-järjestelmä on luotu seuraavalla koodilla:

```
rankhazardplot <- function(...) UseMethod("rankhazardplot")
```

Se kutsuu käytettyjen argumenttien perusteella `rankhazardplot.coxph`-, `rankhazardplot.cph`- tai `rankhazardplot.default`-funktiota, jotka sisältyvät `rankhazard`-pakettiin. Objektit `coxph` ja `cph` sisältävät sovitetun Coxin mallin tulokset. Objektien luominen on käsitelty luvuissa 3.1.1 ja 3.1.2. Mikäli käytössä ei ole `coxph`- tai `cph`-objektia, kutsutaan `rankhazardplot.default`-funktiota.

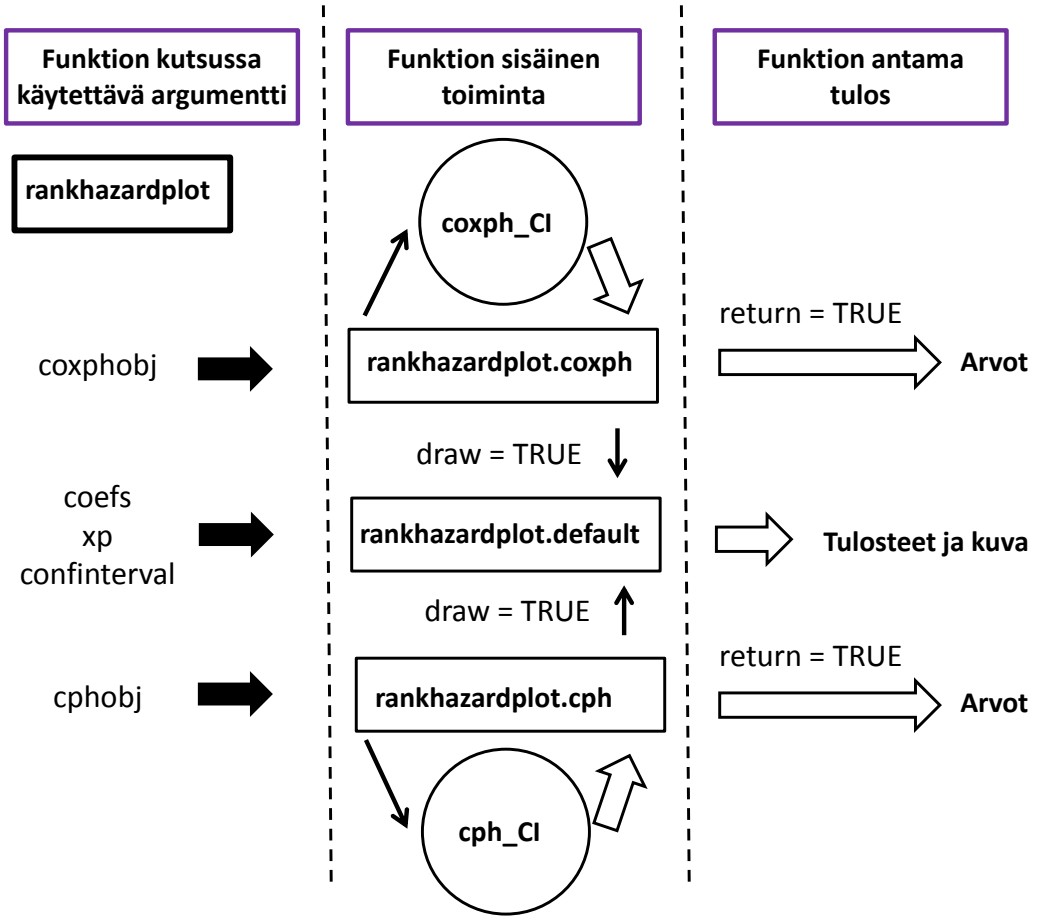
Edellä esitettyjen funktioiden lisäksi `rankhazard`-paketti sisältää kaksi funktiota, joita käytetään luottamusvälien laskemisessa: `coxph_CI` ja `cph_CI`. Käyttäjä ei voi itse kutsua näitä funktioita, vaan funktio `rankhazardplot.coxph` kutsuu funktiota `coxph_CI`. Vastaavasti funktio `rankhazardplot.cph` kutsuu `cph_CI`-funktiota. Paketissa olevien funktioiden sisäinen kutsuntajärjestelmä on havainnollistettu kuvaan 10.

Riskitiheyskuvion piirtäminen tapahtuu `rankhazardplot.default`-funktiolla. Piirtämiseen tarvittavat tiedot voidaan antaa sille suoraan. Tiedot voivat olla joko kovariaattien arvot sekä Coxin mallin kertoimet (argumentti `coefs`) tai funktion `rankhazardplot.coxph` tai `rankhazardplot.cph` palauttamia arvoja (`xp` tai `confinterval`).

Piirtäminen on myös mahdollista käyttäen malliobjektia. Malliobjektin käyttö on suositeltavaa yksinkertaisuuden johdosta. Malliobjektit, joiden luokka on `coxph` tai `cph` sisältävät paljon tietoa sovitetusta Coxin mallista sekä sovittamiseen käytetystä aineistosta. Näitä tietoja käytetään funktioissa `rankhazardplot.coxph` ja `rankhazardplot.cph`, kun koostetaan arvot, joita suhteellisen riskitiheyden jakauman piirtämiseen tarvitaan. Nämä arvot välitetään funktiolle `rankhazardplot.default`, joka laskee suhteelliset riskitiheydet sekä piirtää kuvaajan. Luottamusvälit piirretään, mikäli ne on pyydetty. Lisäksi `rankhazardplot.default`-funktio tekee tulosteet R-ympäristön komentoikkunaan.

6.2 Eri tavat kutsua `rankhazardplot`-funktiota

Käydään seuraavaksi läpi erilaiset tavat kutsua `rankhazardplot`-funktiota sekä syy kunkin tavan käyttämiseen. Kutsuista on esitetty mahdollisimman pel-



Kuva 10: Kaavio, joka näyttää, mitä `rankhazardplot`-funktion kutsumisen jälkeen tapahtuu. Paksuntamatonta tekstiä on käytetty argumenttien nimissä. Paksut mustat nuolet osoittavat, mitä funktiota S3-järjestelmä käyttää kullekin argumentille. Ohuet mustat nuolet osoittavat funktioiden keskinäisen kutsujärjestyksen. Paksut valkoiset nuolet näyttävät, mitä funktio palauttaa. Käyttäjä ei voi kutsua ympyröiden sisällä olevia funktioita.

kistetty muoto, eli vähimmäismäärä argumentteja, joita käyttäen riskitiheyskuvio saadaan piirrettyä. Argumenteista on annettu vain nimi, mikäli arvon esittäminen ei anna lisää informaatiota toiminnasta.

Yksinkertaisin tapa kutsua `rankhazardplot`-funktiota on käyttää malliobjektia. Malliobjekti voi olla luokaltaan joko `coxph` tai `cph`. Vastaavat funktiokutsut ovat:

```
rankhazardplot(coxphobj, data)
rankhazardplot(cphobj, data)
```

Funktio tunnistaa käytetyn objektin luokan automaattisesti, joten käyttäjän näkökulmasta nämä funktiokutsut eivät eroa toisistaan.

Mikäli halutaan yhdistää samaan riskitiheyskuvioon suhteellisen riskitiheyden jakauman kuvaajia eri malleista vertailua varten, ne voidaan piirtää käyttäen kovariaattien arvoja sekä kertoimia:

```
rankhazardplot(x, coefs)
```

Tämä toimii vain numeerisille kovariaateille sekä niiden yksinkertaisille muunnoksille. Mikäli mallissa on esimerkiksi kolmitasoinen faktori, piirtämiseen tarvitaan kertoimien sijasta argumenttia `xp`, joka sisältää termittäiset ennusteet jokaiselle piirrettävälle kovariaatille. Tällöin tulee antaa myös vertailuarvot kaikille piirrettäville kovariaateille. Kutsu on:

```
rankhazardplot(x, xp, refvalues)
```

Kutsussa käytettävät arvot saadaan kutsumalla `rankhazardplot`-funktiota malliobjektia käyttäen, kun annetaan argumentin `return` arvoksi `TRUE`.

Kuvasta 10 nähdään, että oletusfunktio `rankhazardplot.default` piirtää riskitiheyskuvion. Kuvan piirtäminen malliobjektia käyttäen on toteutettu siten, että funktio `rankhazardplot.coxph` tai `rankhazardplot.cph` kutsuu oletusfunktiota `xp`-argumenttia käyttäen. Jos halutaan piirtää luottamusvälit, tällöin käytettävä argumentti on `confinterval`.

Luottamusvälien piirtämiseen on kaksi vaihtoehtoa. Ensimmäinen tapa on käyttää malliobjektia:

```
rankhazardplot(coxphobj, data, confint = TRUE)
```

Tällöin oletuksena piirretään kaikki mallissa olevat kovariaatit ja niiden luottamusvälit. Toinen tapa on käyttää funktion palauttamia arvoja, jolloin oletuksena piirretään vain mallissa ensimmäisenä olevan kovariaatin suhteellisen riskitiheyden jakauma sekä luottamusvälit:

```
arvot <- rankhazardplot(coxphobj, data, return = TRUE)
rankhazardplot(confinterval = arvot$confinterval)
```

Ensimmäinen luottamusvälien piirtämiseen käytetyistä kutsuista on yksinkertaisempi, sillä tarvitaan vain yksi kutsu. Tällöin jälkimmäinen kutsu tapahtuu funktion sisällä. Toinen tapa piirtää luottamusvälit mahdollistaa luottamusvälien piirtämisen samaan kuvaan eri malleista. Tämä ei välttämättä ole erityisen tarpeellinen ominaisuus, sillä useiden luottamusvälien piirtäminen samaan kuvaan voi tehdä kuvista vaikeasti luettavia. Kuvaan piirretään nimittäin jokaista kovariaattia kohden kolme kuvaajaa: suhteellisen riskitiheyden jakauma sekä sen ala- ja yläluottamusväli.

Luottamusvälien piirtämiseen käytetyt kutsut toimivat samalla tavalla myös `cph`-luokan mallille.

6.3 Muutokset versioiden 0.8-1 ja 1.0 välillä

Olen kehittänyt Juha Karvasen tekemää `rankhazard`-pakettia version 0.8-1 pohjalta (Karvanen, 2012). Muokkausten jälkeen julkaistu versio on numeroltaan 1.0 (Karvanen & Koski, 2014). Olen muokannut koodia kahdella tavalla: korjaamalla puutteellisia toimintoja sekä lisäämällä uusia ominaisuuksia.

Tärkein korjaus toimintoihin on faktoreiden ja muunnosten käyttämisen mahdollisuus mallissa, josta riskitiheyskuvio piirretään. Tämä monipuolistaa kuvion käyttömahdollisuuksia merkittävästi. Suurin syy faktoreiden ja muunnosten toimimattomuuteen aikaisemmassa versiossa olivat oletusasetukset sekä kvartiilien laskutapa. Muunnosten käyttäminen olisi onnistunut aiemmassa versiossa, jos funktiolle olisi antanut oletusaineiston tilalle mallissa olevat muuttujat alkuperäisestä aineistosta. Kvartiilien laskutapa taas aiheutti faktoreiden toimimattomuuden. Kvartiilit eivät nimittäin ole määritelty faktoreille, jos tasoilla ei ole järjestystä. Suhteellisen riskitiheyden jakauman piirtämiseksi on kuitenkin mahdollista käyttää mitä tahansa selkeää järjestystä, kuten aakkosjärjestys.

Korjasin koodista muitakin puutteita. Uudessa versiossa vertailukohdan vaihtaminen onnistuu malliobjektia käytettäessä. Lisäksi jakaumia kuvaavien viivojen paksuuden, tyylin ja värin muuttaminen toimii loogisesti ja mahdollisimman yksinkertaisesti, kuten myös kvartiileja ja vaihteluväliä korostavien pisteiden värin ja muodon vaihtaminen.

Kuvan informaatioarvo kärsii, mikäli jakaumia ei nimetä. Siksi lisäsin eri vaihtoehtoja, joista tekstit selitelaatikkoon sekä vaaka-akselille voidaan ottaa, mikäli käyttäjä ei anna niitä funktiolle. Funktion helppokäyttöisyyden säilyttämiseksi halusin, että oletusarvot ovat olemassa. Lisäsin myös virheilmoituksen, jonka funktio antaa, jos kuviotyypin määrittelevä syöte on tuntematon. Muuten saatava virheilmoitus ei olisi helposti yhdistettävissä ongelman lähteeseen.

Uusien ominaisuuksien lisääminen on vaatinut uusien argumenttien määrittelyä `rankhazardplot`-funktiolle. Funktion argumenttien määrä versioiden välillä kasvoi 13:sta 38:aan. Lisätyt argumentit ovat: `data`, `select`, `na.rm`, `draw`, `return`, `CI_level`, `x_CI`, `confint`, `confinterval`, `refpoints`, `ylab`, `ylim`, `yticks`, `yvalues`, `legendlocation`, `axistextposition`, `reftick`, `refline`, `refline.col`, `refline.lwd`, `refline.lty`, `lty`, `bg`, `pt.lwd` ja `cex`. Käydään seuraavaksi läpi uusien argumenttien mahdollistamat ominaisuudet pääpiirteissään. Tarkat kuvaukset argumenttien toiminnasta ovat liitteen A alaluvuissa.

Uudistetulle funktiolle voi antaa alkuperäisen aineiston kokonaan argumentilla `data`, kun piirtämisessä käyttää malliobjektia. Ei siis tarvitse valita itse aineistosta vain mallissa olevia tai piirrettäviä kovariaatteja. Aineiston ei tarvitse olla sama kuin millä malli on sovitettu, kuten ei edellisessäkään versiossa tarvinnut.

Oletuksena riskitiheyskuvioon piirretään kaikkien mallissa olevien muuttujien suhteellisen riskitiheyden jakauman kuvaajat. Tämä voi kuitenkin tehdä

kuvioista raskaslukuisia, mikäli kovariaatteja on paljon. Koska kuvion piirtäminen on helpointa käyttäen malliobjektia, uuteen versioon on lisätty piirrettävien kovariaattien valitsemisominaisuus.

Puuttuvien havaintojen käsittelyä on helpotettu. Edellisessä versiossa funktio käytti kaikkea annettua aineistoa kuvion piirtämisessä. Jos siis halusi piirtämisessä käytettävän vain mallin sovituksessa mukana ollutta aineistoa, funktiolle tuli itse antaa arvot vain mallissa mukana olleista havaintoyksiköistä. Toisaalta funktio käytti oletuksena juuri tällaista aineistoa. Versiossa 1.0 aineistolle ei ole oletusarvoa, vaan se on annettava itse. Jotta aineiston käsittely säilyisi yksinkertaisena, on lisätty argumentti `na.rm`, jota käyttäen voi määritellä, haluaako piirtämisessä käyttää vain täydellisiä, mallissa mukana olevia havaintoja, vai kaikkea saatavilla olevaa informaatiota. Oletuksena vain mallissa olleiden havaintoyksiköiden arvoja käytetään.

Muokattu funktio palauttaa jakaumien piirtämiseen tarvittavat arvot (liitteet A.6 ja B.5), mikäli käyttäjä niitä tarvitsee. Tämä on hyödyllinen ominaisuus, jos haluaa piirtää samaan kuvaan suhteellisen riskitiheyden jakaumia eri malleista. Mikäli tarkoituksena on vain laskea kyseiset arvot, kuvaajan piirtämisen saa myös pois päältä argumenttia `draw` käyttäen.

Uudessa versiossa suhteellisen riskitiheyden jakaumalle on mahdollista piirtää luottamusvälejä. Se onnistuu alkuperäisille muuttujille ja niiden yksinkertaisille muunnoksille sekä faktoreille. Yksinkertainen muunnos on esimerkiksi muuttujan logaritmi. Luottamusvälien laskemisessa käytetään oletuksena mallin sovituksessa käytettyä aineistoa.

Lisäsin funktioon argumentin `refpoints`, joka yksikäsittäistää sen, annetaanko funktiolle suhteellisen riskitiheyden jakaumien vertailukohdat vai vertailuarvot. Ero käsitteiden välillä on määritelty taulukossa 2 (sivu 17). On myös mahdollista vaihtaa vain osa vertailukohdista antamalla muille kovariaateille vertailukohdaksi `NA`. Näille muuttujille vertailukohtana käytetään oletusarvoa.

Graafisten ominaisuuksien parantaminen oli keskeistä kehitysprosessissa, jotta kuvioista voi tehdä monipuolisia. Versioon 1.0 on tullut useita uusia graafisia ominaisuuksia. Edellisessä versiossa pystyakselia ei voinut muokata ollenkaan. Nyt pystyakselille voi vaihtaa otsikon, vaihteluvälin, asteikon merkkien paikat ja asteikolla näkyvät luvut. Uudistukset ovat merkityksellisiä ainakin, kun kuvia halutaan verrata keskenään.

Nyt käyttäjä voi itse määrätä selitelaatikon sijainnin. Selitelaatikon voi myös jättää piirtämättä, ja lisätä sen jälkikäteen, jolloin esimerkiksi tekstien kokoa pääsee muuttamaan. Vaaka-akselin selitetekstien sisennystä voi myös vaihtaa, jos selitteet esimerkiksi menevät päällekkäin muiden tekstien kanssa.

Myös referenssin näkyminen pystyakselilla selkeyttää kuvaa. Se on vaikeaa saada oletusarvoisesti näkyviin ilman, että referenssi menee päällekkäin muiden asteikon lukujen kanssa, sillä arvot piirretään alhaalta ylöspäin. Referenssin korostamiseksi on myös kaksi lisättyä keinoa: referenssin asteikkomerkin lihavointi sekä referenssisuoran lisääminen. Referenssisuora on vaakasuora re-

Taulukko 9: Vertailu `rankhazard`-paketin sisältämien funktioiden koodien rivimääristä versioiden 0.8-1 ja 1.0 välillä. Tyhjät rivit ja pelkkiä kommentteja sisältävät rivit on jätetty pois laskuista. Versiossa 0.8-1 ei ollut funktioita `coxph_CI` ja `cph_CI`.

Funktio	Versio 0.8-1	Versio 1.0
`rankhazardplot`	1	1
`rankhazardplot.default`	82	170
`rankhazardplot.coxph`	31	95
`rankhazardplot.cph`	21	86
`coxph_CI`	-	38
`cph_CI`	-	38
yhteensä	135	428

ferenssin kohdalla ja sen väriä, tyyliä ja paksuutta voi muuttaa. Oletuksena se piirretään mustana katkoviivana. Referenssisuora piirretään muiden kuvaajien alle.

Uudistetussa versiossa jokaisen jakaumaa kuvaavan viivan tyyliä voi muuttaa erikseen. Pisteiden värin ja muodon muuttamisen lisäksi voi muuttaa pisteiden kokoa sekä käyttää pistetyyppejä, joille voi piirtää eri värisen sisuksen kuin reunan.

Lisäksi olen kehittänyt `rankhazardplot`-funktion dokumentaatiota (liite B), ja tehnyt siihen paljon esimerkkejä. Liitteessä C on myös tulosteet kaikista esimerkeistä. Niistä nähdään uusi ominaisuus: R-ympäristön komentoikkunaan tulostuvat taulukot. Ne näyttävät suhteellisen riskitiheyden jakaumien arvojen vaihteluvälit ja kvartiilit kovariaateittain.

Paketin ominaisuuksien monipuolistamisen myötä koodin määrä lisääntyi 293 riviä. Tarkempi erittely on koottu taulukkoon 9. Version 1.0 koodit ovat kokonaisuudessaan liitteessä D.

7 Pohdinta

Riskitiheyskuvio täyttää Tuften (1983) kirjaamat laadukkaan graafisen esityksen kriteerit: se on selkeä ja esittää mallin tulokset pienessä tilassa sekä sisältää paljon informaatiota. Useiden kovariaattien piirtäminen samaan kuvaan mahdollistaa sekä kovariaattien vertailun että optimaalisen tilankäytön. Riskitiheyskuvio esittää jokaisen kovariaatin suhteellisen riskitiheyden jakauman populaatioestimaatin, jonka tulkinta on looginen: vaaka-akseli ottaa huomioon arvojen yleisyyden, ja mitä korkeammalla kuvaaja kulkee, sitä suurempi kovariaattiin liittyvä suhteellinen riskitiheys on. Riskitiheyskuviosta katsoja saa kuvaa myös mallintamisessa käytetystä aineistosta.

Tutustuttuani muihin Coxin mallin diagnostiikkakeinoihin olen huomannut niiden keskittyvän mallin sopivuuden ja suhteellisen riskitiheyden oletuksen tutkimiseen, usein erilaisten residuaalien kautta. Mallin sopivuuden ja oletusten tutkiminen on toki tärkeää mallintamisen kannalta, mutta ne eivät kerro katsojalle erityisen paljon käsiteltävän mallin sisällöllisestä tulkinnasta. Riskitiheyskuvio on erilainen, sillä sitä käyttäen saadaan esitettyä sekä aineistoa että mallin tuloksia. Riskitiheyskuvio tarjoaa helposti tulkittavan tavan tarkastella mallin mielekkyyttä sekä kovariaatteihin liittyviä riskitiheyksiä populaatiotasolla. Mielestäni riskitiheyskuvio on erittäin tervetullut lisä Coxin mallin diagnostiikkakeinoihin.

Pro gradu -tutkielmani tavoitteena on ollut parantaa `rankhazard`-pakettia. Paketti ja funktiot on tehty R-ympäristön vaatimusten mukaan. Tekemäni versio on osa R-ohjelmapakettien kokoelmaa, CRANia, ja se löytyy osoitteesta `http://CRAN.R-project.org/package=rankhazard`. Olen täydentänyt myös `rankhazardplot`-funktion dokumentaation R-ympäristön dokumentaatiotyyliä noudattaen. Liitteessä B oleva teksti on julkaistu Internetiin osoitteeseen `http://cran.r-project.org/web/packages/rankhazard/rankhazard.pdf`. Tämän tutkielman myötä funktiolle on tarjolla vielä kattavampi kommentointi suomeksi. Tutkielmani tarjoaa useita esimerkkejä sekä funktion käytöstä ja toiminnasta että riskitiheyskuvion tulkinnasta. Olen erittäin ylpeä siitä, että olen pro gradu -työnäni kehittänyt ohjelmistoa, joka on jaettu käyttöön kaikille ympäri maailmaa. Uskon tämän olevan melko poikkeuksellista opintojen päättötyölle.

Paketin `rankhazard` versio 1.0 ei kuitenkaan ole täydellinen, mikä on ohjelmistoille tyypillistä. Olen löytänyt julkaisun jälkeen koodista yhden koodausvirheen, joka liittyy vertailukohtien muuttamiseen ja ilmenee vain erikoistapauksessa. Paketin funktiot kaipaavat myös lisää virheilmoituksia. Esimerkiksi, kun funktiota kutsutaan käyttäen `xp`-argumenttia, tulisi testata, että aineiston ja termittäisten ennusteiden koko on sama. Tällä hetkellä funktio ei anna virheilmoitusta ja piirtää kuvan väärin. Tämä voi aiheuttaa päänvaivaa käyttäjälle. Lisäksi koodia voisi selkeyttää järjestämällä osia eri tavalla.

Vasta version1.0 julkaisemisen jälkeen löysin R-ympäristöstä `matplot`-funktion. Sen käyttäminen voisi yksinkertaistaa kirjoittamaani koodia ja tuoda

funktion käyttäjille lisää mahdollisuuksia kuvan muokkaamiseen, vaikka graafiset argumentit ovat tässäkin versiossa helppokäyttöiset ja loogiset. Tällä hetkellä esimerkiksi viivojen värien ja tyylien vaihtaminen on vaatinut `rankhazardplot`-funktiolle useita lisäargumentteja, jotka mitä luultavimmin saisi liitettyä argumenttiin '...', mikäli piirtäminen tapahtuisi `matplot`-funktiota käyttäen. Tällöin `rankhazardplot`-funktiolle saataisiin automaattisesti käyttöön myös argumentti `add`, joka helpottaisi eri malleista piirrettävien jakaumien yhdistämistä samaan riskitiheyskuvioon.

Versiossa 1.0 luottamusvälejä ei voida laskea kaikille muuttujille, mikä on hieman harmillista. Jakauman luottamusväli saadaan käyttämällä suhteellisen riskitiheyden laskemisessa kertoimen estimaatin sijaan sen luottamusväliä. Faktorin tapauksessa luottamusväli saadaan riskitiheyssuhteiden luottamusvälinä. Monimutkaisimmille muunnoksille sama ei onnistu, sillä termittäistä ennustetta ei voida laskea suoraan kovariaatin arvon ja kertoimen tulona. Luottamusvälien toteuttaminen termittäisten ennusteiden keskivirheiden avulla saattaisi olla mahdollista ja olisi myös jatkotutkimuksen arvoista.

Haluan jakaa hieman kokemustani ohjelmakoodin kirjoittamisesta, etenkin korjaamisesta. Kun korjaa valmista koodia, itsensä tai jonkun muun tekemää, kannattaa ihan ensimmäiseksi selvittää, miksi koodi ei toimi toivotulla tavalla. Siitäkin huolimatta, että itsellään olisi tiedossa jokin muu toteutusvaihtoehto. Sain tämän prosessin aikana huomata, että olisi ollut paljon helpompaa korjata alkuperäinen koodi toimivaksi kuin kirjoittaa täysin uusi. Toki tämä keksimäni "uusi tapa" pääsi käyttöön luottamusvälien laskemisessa. Suosittelen myös heti alussa selvittämään, millä eri tavoin funktion tulisi toimia, jotta voidaan löytää paras mahdollinen tapa toteuttaa ne kaikki.

Aion tehdä `rankhazard`-pakettiin päivityksen. Ensisijaiset muutokset koskevat funktion käytettävyyttä: kuvaajien lisäämistä riskitiheyskuvioon `add`-argumenttia käyttäen sekä virheilmoitusten monipuolistamista. Jo `matplot`-funktion käyttämiseen siirtyminen korjannee lisäämisominaisuuden lisäksi osan puuttuvista virheilmoituksista. Se tekisi koodista myös todennäköisesti lyhyemmän ja selkeämmän. Mikäli jostain syystä `matplot`-funktion käyttäminen ei ole mahdollista, voi lisäämisominaisuuden toteuttaa muutenkin.

Kuitenkin versio 1.0 on jo erittäin toimiva ja monipuolinen. Päivittämäni muutokset ovat parantaneet riskitiheyskuvion käyttömahdollisuuksia merkittävästi. Nyt suhteellisen riskitiheyden jakaumia voi piirtää monenlaisista muuttujista, mukaan lukien muunnokset ja faktorit. Kuvan graafiset ominaisuudet ovat käyttäjän muokattavissa. Myös luottamusvälien piirtäminen on mahdollista useille muuttujatyypeille. Lisäsin käyttöä helpottavia ominaisuuksia, kuten muuttujien osavalinnan. Funktion käyttäjä voi piirtää haluamansalaisen kuvan melko helposti. Mielestäni riskitiheyskuviolla on nyt hyvät mahdollisuudet yleistyä Coxin mallin diagnostiikkakeinojen joukossa. Erityisen iloinen olen siitä, että ideoimaani tapaa riskitiheyskuvion käyttämiseen ajassa muuttuvilla kovariaateilla ovat jo käyttäneet artikkelissaan Reinikainen, Laatikainen, Karvanen ja Tolonen (2014).

A Rankhazardplot-funktion argumentit ja toiminta

A.1 Funktion `rankhazardplot` toiminnalliset argumentit

Tässä liitteessä sekä liitteissä A.2 ja A.6.1 esitellään funktion `rankhazardplot` argumentit. Argumentit kuuluvat pääasiassa kahteen luokkaan: toiminnallisiin ja graafisiin. Lisäksi on funktion palauttamiin arvoihin sekä kuvan piirtämiseen liittyvät argumentit `return` ja `draw`. Toiminnalliset argumentit ovat laskentaan ja piirtämiseen liittyviä, graafiset kuvan ulkonäköön vaikuttavia. Näiden välillä on toinenkin ero: graafisia argumentteja voi käyttää minkä tahansa `rankhazardplot`-funktion kutsuntatavan kanssa, mutta toiminnallisia ei. Taulukkoon 10 on koottu eri tavat kutsua `rankhazardplot`-funktiota sekä tapojen kanssa käytettävät argumentit. Tekstissä viitataan litteessä D oleviin koodeihin rivinumeroilla, jotta lukijan on helpompi tarkastella funktion toimintaa. Esimerkit ovat taas peräisin funktion dokumentaatiosta, ja ne löytyvät liitteestä C. Argumenttien esittelyä syventää myös `rankhazardplot`-funktion dokumentaatiossa oleva argumenttien esittely, joka on liitteessä B.3.

Taulukko 10: Funktion `rankhazardplot` kutsuntatavan määräävä argumentti sarakkeella (**lihavoituna**) sekä sen kanssa pakolliset argumentit (P), valinnaisena annettavat argumentit (niiden oletusarvo), ja argumentit, joita ei voi käyttää (–). Taulukosta nähdään esimerkiksi se, että argumenttia `xp` käytettäessä tulee aineisto antaa argumenttina `x` ja vertailuarvojen antaminen argumenttina `refvalues` on pakollista. Graafiset argumentit on jätetty tarkastelun ulkopuolelle, sillä niiden käyttö ei muutu kutsuntavan mukaan.

	coxphobj	**cphobj**	**coefs**	**xp**	**confinterval**
`data`	P	P	–	–	–
`x`	–	–	P	P	–
`select`	kaikki	kaikki	–	–	1
`refpoints`	mediaanit	mediaanit	mediaanit	–	–
`refvalues`	–	–	–	P	–
`CI_level`	0.95	0.95	–	–	–
`x_CI`	ks. A.1.7	ks. A.1.7	–	–	–
`confint`	`FALSE`	`FALSE`	–	–	–
`na.rm`	`TRUE`	`TRUE`	`TRUE`	`TRUE`	`TRUE`
`draw`	`TRUE`	`TRUE`	–	–	–
`return`	`FALSE`	`FALSE`	–	–	–

A.1.1 Coxin malli: `coxphobj` tai `cphobj`

Funktiota `rankhazardplot` voidaan kutsua antamalla `coxph`-luokan objekti argumenttina `coxphobj` tai `cph`-luokan objekti argumenttina `cphobj`. Malliobjektin käyttö tekee `rankhazardplot`-funktion kutsumisesta yksinkertaista, sillä yhdellä argumentilla saadaan annettua paljon informaatiota. Malliobjektia käyttämällä ennusteiden laskeminen onnistuu `survival-` ja `rms-` paketeista löytyvien `predict.coxph-` ja `predict.rms-` funktioilla (Liite D.3 rivit 274, 276; liite D.4 rivit 394–395). Malliobjekti kertoo, mitä kovariaatteja malli sisältää, ja ne voidaan valita aineistosta, joka sisältää muitakin muuttujia (Liite D.3 rivit 228, 233–241; liite D.4 rivit 355, 360–361). Malliobjektista selviää myös, mitkä muuttujista ovat faktoreita. Tämä tieto on tärkeä, sillä faktoreiden ja numeeristen muuttujien käsittely poikkeaa toisistaan (Liite D.3 rivit 245–248; liite D.4 rivit 365–368).

Malliobjektit tulee luoda käyttämällä asetusta `x = TRUE`. Tällöin malliobjekti sisältää mallin sovituksessa käytetyn aineiston nimellä `x`. Muuten argumentti `x_CI` tulee antaa, jotta luottamusvälit voidaan laskea. Tämän aineiston tyyppi poikkeaa tavallisesta aineistosta siten, että esimerkiksi faktorit esitetään indikaattorimuuttujina. (Liite D.3 rivit 209, 221–226; liite D.4 rivit 336, 348–353.) Luottamusvälit lasketaan tätä aineistoa käyttäen. Laskemiseen tarvitaan myös mallin avulla saatavat Coxin mallin kertoimien luottamusvälit (Liite D.3 rivit 285–286; liite D.4 rivit 404–405). Malliobjektia käytetään myös funktioissa `coxph_CI` ja `cph_CI`, jotka palauttavat luottamusvälien laskemiseen tarvittavat termittäiset ennusteet (Liite D.3 rivit 291–293; liite D.4 rivit 410–412). Malliobjektin avulla tunnistetaan faktorit, niiden tasot ja indikaattorimuuttujien sijainti aineistossa, ja niiden perusteella voidaan luoda alkuperäinen faktori (Liite D.5 rivit 457, 466, 471, 474, 478; liite D.6 rivit 523, 532, 537, 540, 544).

Esimerkkikoodi malliobjektin käyttämisestä kuvan piirtämisessä on liitteessä C.1. Kuva 11 sisältää kaksi kertaa saman kuvan: sekä `coxph-` että `cph`-malliobjektilla piirrettynä.

A.1.2 Kovariaattien arvojen antaminen: `data` tai `x`

Kovariaattien arvot annetaan kutsutavasta riippuen joko argumenttina `data` tai `x`. Ero näissä argumenteissa on se, että `data` voi sisältää muitakin kovariaatteja kuin mallissa on käytetty eikä kovariaattien järjestyksellä ole väliä. Kuitenkin jokaista mallissa olevaa muuttujaa vastaava kovariaatti on löydyttävä aineistosta, muuten ennusteiden laskeminen ei onnistu. Datamatriisista `data` valitaan käsiteltäväksi aineistoksi vain mallissa olevat muuttujat, ja ne järjestetään vastaavasti kuin mallin sovituksessa on määritelty. (Liite D.3, rivit 209, 218–219, 233, 241; liite D.4, rivit 336, 345–346, 361.)

Datamatriisin `x` taas tulee sisältää valmiiksi vain piirrettävät kovariaatit ja juuri siinä järjestyksessä, jossa jakaumat piirretään. Järjestyksen tulee olla sama kuin argumentissa `xp` tai `coefs`, joita käytetään yhtä aikaa tämän

aineistotyypin kanssa. (Liite D.2, rivit 2, 21, 25–26, 36, 43–44, 133, 164.)

A.1.3 Piirrettävien kovariaattien valinta: `select`

Kun Coxin malli sovitetaan, kovariaatit kirjoitetaan kutsuun jossain järjestyksessä. Tämä järjestys määrää numeron, jolla kovariaattiin viitataan. Argumentilla `select` voidaan tehdä osavalinta ja määrittää vektoria käyttäen kovariaatit, joiden suhteellisen riskitiheyden jakaumat piirretään kuvaan. Piirrettävien kuvaajien järjestys määräytyy siten, että vektorin ensimmäisessä indeksissä oleva numero määrää ensimmäisenä piirrettävän kuvaajan. Siten päällimmäinen kuvaaja piirretään kovariaatille, johon viitataan viimeisessä indeksissä.

Argumenttia `select` voidaan käyttää joko malliobjektin tai `confinterval`-argumentin kanssa. Jos kutsuargumentti on `xp` tai `coefs`, valinta tulee tehdä ennen `rankhazardplot`-funktion kutsumista. Malliobjektia käytettäessä piirretään oletuksena kaikkien mallissa olevien muuttujien suhteellisen riskitiheyden jakaumat ja `confinterval`-argumentin kanssa vain ensimmäisen. Tässä tulee huomata, että jos luottamusvälit piirretään suoraan malliobjektikutsulla, on oletusarvona kaikki jakaumat. (Liite D.2 3, 14, 17–18, 75, 77; liite D.3, rivit 209, 230–231, 261–271, 322; liite D.4, rivit 336, 357–358, 382–392, 444).

Esimerkkikuva 14 sivulla 82 osoittaa sen, ettei kaikkia mallin kovariaattien suhteellisen riskitiheyden jakaumia kannata aina piirtää samaan kuvaan. Mikäli kovariaatteja on paljon, ja ne kaikki halutaan esittää, tulokset voi jakaa useampaan kuvaan.

A.1.4 Coxin mallin kertoimet: `coefs`

Riskitiheyskuvio voidaan piirtää Coxin mallin kertoimia käyttäen, mikäli piirrettävät kovariaatit ovat numeerisia. Toinen ehto on, että muuttujalle on estimoitu mallissa vain yksi kerroin. Näille muuttujille voidaan laskea termittäiset ennusteet kertoimen ja havaintojen tulona. Näitä termittäisiä ennusteita käytetään suhteellisen riskitiheyden laskemisessa.

Coxin mallin kertoimista valitaan argumenttiin `coefs` vain piirrettävien kovariaattien kertoimet. Muuttujien arvot annetaan samalla periaatteella vastaavassa järjestyksessä datamatriisina argumentille `x`. Mikäli kovariaatille on tehty yksinkertainen muunnos, sama muunnos tulee tehdä kovariaatin arvoille, joita käytetään laskemisessa. Tällöin vaaka-akselille tulostettavat (liite A.4.2) tunnusluvut lasketaan muunnetuista arvoista. (Liite D.2, rivit 2, 30–31, 35–44.)

Argumentilla `coefs` on yksinkertaista piirtää suhteellisen riskitiheyden jakaumia eri malleista samaan kuvaan, kuten liitteessä C.9 on tehty. Toisessa esimerkissä sivulta C.3 on näytetty, kuinka kertoimia käyttäen kuvan voi piirtää sekä alkuperäisellä että `rankhazardplot`-funktion palauttamalla aineistolla.

A.1.5 Coxin mallin ennusteet termeittäin: `xp`

Riskitiheyskuvio voidaan piirtää myös antamalla funktiolle suoraan termittäiset ennusteet argumentilla `xp`. Käytetäänpä funktiolle `rankhazardplot` mitä tahansa kutsuntatapaa, arvot muunnetaan funktion sisällä ensin termittäisiksi ennusteisiksi. Liitteessä A.4.4 käsitellään sitä, kuinka termittäisiä ennusteita käytetään suhteellisen riskitiheyden laskemisessa. Haluttaessa funktio `rankhazardplot` palauttaa termittäiset ennusteet, joita voidaan käyttää eri mallien kovariaattien suhteellisten riskitiheyksien jakaumien piirtämisessä samaan riskitiheyskuvioon. Tällöin tulee antaa myös suhteellisen riskitiheyden vertailuarvot, sillä niitä ei voida määrittää muuten annettavien argumenttien avulla. (Liite D.2, rivit 2, 23, 30–33.)

Argumenttia `xp` käyttäen voi itse päättää, haluaako esimerkiksi logaritmimuunnetulle kovariaatille vaaka-akselin arvoiksi muunnetut vai alkuperäiset arvot. Tällöin nimittäin datamatriisin `x` arvoja käytetään vain vaaka-akselin tunnuslukujen määrittämiseen, ei lainkaan suhteellisen riskitiheyden laskemiseen. Malliobjektia käytettäessä arvot vaaka-akselilla ovat aina alkuperäiset ja kertoimia käytettäessä muunnetut.

Esimerkissä liitteessä C.2 on esitetty kaikkien mallissa olevien kovariaattien piirtäminen `xp`-argumenttia käyttäen. Liitteessä C.7 ja C.8 kuvissa 17 ja 18 on esitelty `xp`-argumentin käyttöä. Esimerkeissä yhdistetään eri malleista peräisin olevia kuvaajia samaan riskitiheyskuvioon.

A.1.6 Suhteellisen riskitiheyden vertailukohdan tai -arvon muuttaminen: `refpoints` ja `refvalues`

Vertailukohdan oletusarvon määrittämistä ja muuttamista `rankhazardplot`-funktiossa käsitellään liitteessä A.4.3. Oletuksena vertailukohta on numeerisille kovariaateille mediaani ja faktorille vertailutaso. Vertailukohdat voidaan vaihtaa antamalla vertailukohdat tai -arvot vektorissa, joka sisältää arvon jokaiselle piirrettävälle jakaumalle. Arvojen tulee olla myös vastaavassa järjestyksessä kuin `select`-argumentti osoittaa.

Kun vertailukohtana on mediaani, nähdään kuvaajasta riskitiheys verrattuna mediaanilla laskettuun riskitiheyteen. Tällöin kuvasta voidaan katsoa, onko suhteellinen riskitiheys kovariaatin arvojen ääripäissä merkillisen suuri tai pieni verrattuna mediaaniin. Jos taas käytössä on faktori, jonka tasot kuvaavat sairauden etenemistä, voi olla järkevää valita vertailukohta normaalitasoksi eli tilaan, jossa kunto on mahdollisimman hyvä.

Esimerkkikuvassa 19 on muutettu bilirubiinimuuttujan vertailukohdaksi 1.2, jota voidaan pitää normaalin bilirubiiniarvon ylärajana (Clinical Reference Laboratory, Inc.). Vastaavasti logaritmimuunnetulle bilirubiinille vertailuarvo on $\log(1.2)$.

Kuvassa 20 on esitetty vertailukohdan vaihtaminen faktorille, mikä toimii niin, että vertailutason nimi annetaan vektoriin `refpoints` merkkijonona. Sukupuolijakaumasta käy selvästi ilmi, että sairastuneista noin 90 % on naisia.

Sukupuoli ei ole mallissa tilastollisesti merkitsevä muuttuja, vaan se on haluttu esimerkkiin juuri siksi, että se on merkkijonofaktori.

Vertailukohdat on mahdollista vaihtaa myös vain osalle kovariaateista. Tällöin muille annetaan arvoksi `NA`, ja funktio `rankhazardplot` käyttää tällöin oletusarvoja niille kovariaateille. Esimerkissä C.11 havainnollistetaan tätä ominaisuutta sekä vertailukohtien vaihtamista, mikäli käytössä on `xp`-argumentti.

A.1.7 Luottamusvälien laskeminen: `confinterval`, `CI_level`, `confint` ja `x_CI`

Suhteellisen riskitiheyden luottamusvälit lasketaan aina, kun `rankhazardplot`-funktiota kutsutaan malliobjektia käyttäen. Argumentille `x_CI` tulee siksi antaa luottamusvälejä laskettaessa käytettävä aineisto, mikäli malliobjekti ei sisällä sitä (ks. liite A.1.1). Annetun aineiston muotoon ja luottamusvälien laskemiseen perehdytään tarkemmin liitteessä A.4.5.

Oletuksena lasketaan 95 prosentin luottamusväli. Sitä voi muuttaa antamalla malliobjektikutsussa argumentin `CI_level` arvoksi luottamusvälin varmuustason lukuna nollasta yhteen.

Mikäli suhteellisen riskitiheyden jakaumalle haluaa piirtää luottamusvälit, ne tulee pyytää käyttäen argumenttia `confint` tai `confinterval`. Ensimmäinen argumenteista on looginen muuttuja, jota käytetään malliobjektin kanssa. Jos `confint = TRUE` piirretään valituille muuttujille myös luottamusvälit suhteelliselle riskitiheydelle. Toinen tapa piirtää on pyytää funktiota palauttamaan arvoja, joihin kuuluu lista `confinterval`, ja antaa tämä lista samannimiselle argumentille.

Sivulta 95 alkavassa esimerkkikoodissa on esitelty molemmat tavat piirtää luottamusvälit. Kuvista huomataan vertailukohdan vaihtamisen vaikutus luottamusväleihin. Kuvat on piirretty eri malleista, joten vaaka-akselille tulostuvat tunnusluvut eivät ole täsmälleen samat.

A.1.8 Puuttuvien havaintojen käsittely: `na.rm`

Mallinnuksessa käytettävässä aineistossa on usein puuttuvia havaintoja. Mallin sovituksessa käytetään kuitenkin vain niitä havaintoja, joista on mitattu kaikki arvot käytetyistä muuttujista. Näitä havaintoja kutsutaan täydellisiksi. Riskitiheyskuvion piirtämiseen on siis kaksi vaihtoehtoa: joko käytetään vain täydellisiä havaintoja tai kaikkia mitattuja arvoja. Mikäli annetusta aineistosta halutaan käyttää vain täydellisiä havaintoja, annetaan argumentin `na.rm` arvoksi `TRUE`, muuten `FALSE`. Oletusarvo on `TRUE`.

Malliobjektia käytettäessä funktion sisällä valitaan annetusta aineistosta ensin mallissa olevat muuttujat. Tämän jälkeen valitaan täydelliset havainnot, mikäli näin on määritetty. Näin siksi, että aineistoa karsitaan vain mallissa olevien muuttujien perusteella. Valittua aineistoa käytetään laskuissa sekä funktion palauttamissa arvoissa. Tästä valitusta aineistosta välitetään

vielä `select`-argumentin määräämät kovariaatit `rankhazardplot.default`-funktiolle piirtämistä varten. (Liite D.3, rivit 214, 241, 243; liite D.4, rivit 341, 361, 363.)

Jos argumenttina on `confinterval`, piirrettävä aineisto on alunperin annettu argumentissa `x_CI`, jossa on vain mallissa olevat muuttujat. Argumentille `x_CI` on oletusaineisto, joka sisältää vain täydelliset havainnot. Argumentin `na.rm` arvolla on merkitystä vain, jos aineiston on antanut itse. Havaintojen valinta aineistosta tehdään ennen piirrettävien kovariaattien valintaa. Muuten yhden kovariaatin suhteellista riskitiheyden jakaumaa piirrettäessä `na.rm`-argumentin arvolla ei olisi merkitystä. Mikäli mallissa on muuttuja, jolle luottamusvälejä ei voida laskea, on mahdollista, että piirrettävien havaintojen joukkoon jää epätäydellinen havainto, vaikka `na.rm` olisi `TRUE`. (Liite D.2, rivit 7, 13, 16.)

Argumentteja `coefs` tai `xp` käytettäessä piirrettävät muuttujat valitaan jo ennen funktion kutsua. Kertoimia `coefs` käytettäessä havaintojen valinta tehdään datamatriisille `x`, ja termittäiset ennusteet lasketaan tämän jälkeen. Termittäisiä ennusteita `xp` käytettäessä valitaan piirrettävät havainnot molemmista matriiseista `x` ja `xp`. (Liite D.2, rivit 7, 21, 23.)

Kun käytetään malliobjektia, `rankhazardplot.coxph`- tai `rankhazardplot.cph`-funktio kutsuu `rankhazardplot.default`-funktiota. Tällöin argumenttien `x` ja `xp` sisältämät datamatriisit eivät muutu `na.rm`-argumentin arvon myötä enää `rankhazardplot.default`-funktiossa.

Kun piirtämisessä käytettävät havainnot on valittu, määritetään havaintojen lukumäärä. Ennen piirtämistä selvitetään, kuinka monta puuttuvaa havaintoa kullakin kovariaatilla on. Jos käytetään täydellisiä havaintoja, puuttuvia havaintoja ei ole. (Liite D.2, rivit 25, 133.)

Piirtäminen tapahtuu silmukan sisällä, jossa ensin lasketaan piirrettävän kovariaatin havaittujen havaintojen määrä. Tätä lukua käyttäen määritetään skaalatut järjestysluvut sekä kvartiilien indeksit. Suhteellisen riskitiheyden jakauman kuvaajaa sekä sen luottamusvälejä piirrettäessä käytetään vain havaittuja arvoja kovariaatin arvojen suuruusjärjestyksessä.(Liite D.2, rivit 122–125, 136–138, 160–161, 171–172, 174–175.)

A.2 Funktion `rankhazardplot` graafiset argumentit

Graafiset argumentit vaikuttavat kuvan ulkonäköön. Niitä voi käyttää minkä tahansa luvussa 6.2 esitellyn `rankhazardplot`-funktion kutsuntatavan kanssa. Kuva piirretään `rankhazardplot.default`-funktiolla, ja vain se käyttää pääasiassa graafisia argumentteja. Poikkeuksena ovat argumentit `axistext` ja `legendtext`, joille otetaan oletusarvot malliobjektista, jos se on käytössä. Kaikkien muiden argumenttien tapauksessa funktiot `rankhazardplot.coxph` ja `rankhazardplot.cph` välittävät saamansa graafisten argumenttien arvot `rankhazardplot.default`-funktiolle.

Liitteessä C.5 on esimerkkikuva 15, jossa on käytetty useita graafisia ar-

gumentteja. Se toimii esimerkkinä esiteltävien argumenttien käytöstä, mikäli toista esimerkkiä ei osoiteta.

A.2.1 Pystyakselin asteikon valitseminen: `plottype`

Argumentille `plottype` voi antaa merkkijonon ”hazard” tai ”loghazard”. Oletusarvo on ”hazard”. Tällöin piirretään suhteellinen riskitiheys logaritmiselle pystyakselille, ja pystyakselin otsikko on oletuksena ”relative hazard”. Pystyakselin asteikkomerkit määritetään siten, että referenssi 1 kuuluu joukkoon. (Liite D.2, rivit 7, 79–85, 99–105.)

Jos valinta on ”loghazard”, piirretään suhteellisen riskitiheyden logaritmi lineaariselle asteikolle. Tällöin pystyakselin otsikon oletusarvo on ”logarithm of the relative hazard”. Pystyakselin asteikkomerkit määritetään siten, että referenssi 0 kuuluu joukkoon. (Liite D.2, rivit 106–112.)

Pystyakselin otsikkoa voi molemmissa tapauksissa muuttaa `ylab`-argumenttia käyttäen, ja asteikkomerkkien paikkoja voi vaihtaa `yticks`-argumentilla. Mikäli `plottype`-argumentille antaa minkä tahansa muun arvon kuin ”hazard” tai ”loghazard”, funktio palauttaa virheilmoituksen ”Unknown plottype” (Liite D.2, rivit 28–29).

Liitteessä C.6 on kuva 16, jossa mallin kovariaateille on esitetty sekä suhteellisen riskitiheyden jakauma sekä sen logaritmi.

A.2.2 Pystyakselin muokkaaminen: `ylab`, `ylim`, `yticks`, `yvalues`

Pystyakselille voi antaa minkä tahansa otsikon `ylab`-argumenttia käyttäen. Oletusarvo argumentille on selostettu liitteessä A.2.1. Suomenkieliset vastineet oletuksena oleville otsikoille ovat ”suhteellinen riskitiheys” ja ”suhteellisen riskitiheyden logaritmi”. (Liite D.2, rivit 6, 100, 107, 141–144.)

Pystyakselin vaihteluvälin voi muuttaa argumentilla `ylim`. Se on tarpeellista, mikäli halutaan vertailla eri riskitiheyskuvioita keskenään. Tällöin kuvia on helpompi tulkita, jos vaihteluväli on kuvissa sama. Oletuksena vaihteluväli on aineistosta lasketun suhteellisen riskitiheyden (logaritmin) minimi ja maksimi. Mikäli myös luottamusvälit piirretään, niiden arvot otetaan huomioon vaihteluväliä määritettäessä. (Liite D.2, rivit 6, 86–97, 141–144.)

Asteikkomerkkien paikkojen muuttaminen on tarpeellista eri kuvia verratessa, sillä asteikkojen olisi syytä olla samanlaiset. Lisäksi merkkejä voi tihentää tai harventaa. Argumenttia `yticks` käytetään asteikkomerkkien paikkojen määräämiseen. Oletuksena asteikkomerkit määritetään referenssin molemmin puolin paloittain `pretty`-funktioa käyttäen. Se tekee asteikon merkit siten, että numeerinen erotus asteikkomerkkien välillä pysyy samana. Tällöin asteikkoa on helppo lukea, vaikka jokaiselle merkille ei mahtuisi merkitsemään arvoa. Myös referenssin kohtaan piirretään aina asteikkomerkki. (Liite D.2, rivit 6, 101–102, 108–109, 147.)

R-ympäristön `plot`-funktio piirtää asteikon merkit alhaalta ylös. Mikäli merkkien väliin ei jää riittävästi tilaa, merkin selite jätetään piirtämättä. Täl-

lä tavoin kuvasta jää pois helposti juuri referenssin selite, joten käyttäjälle on tarpeellista kyetä itse määräämään kuvaan piirrettävät selitteet. Tällöin voi tietoisesti tehdä tilaa esimerkiksi juuri referenssin selitteelle. Asteikkojen selitteet voidaan määrittää `yvalues`-argumenttia käyttäen. Mikäli arvoja ei anneta, käytetään `yticks`-argumentin sisältämiä arvoja. Kuvan luettavuuden kannalta on tärkeää, että `yvalues` sisältää vain arvoja, joille piirretään myös asteikkomerkki. (Liite D.2, rivit 7, 114, 148)

A.2.3 Vaaka-akselin tekstit ja sisennys: `axistext` ja `axistextposition`

Vaaka-akselille tulostetaan jokaisen piirrettävän kovariaatin vaihteluväli sekä kvartiilit suuruusjärjestyksessä. Akselin alkuun tulostetaan myös kovariaatteja kuvaavat nimet, joita voi vaihtaa `axistext`-argumenttia käyttäen. Nimet tulee antaa vain piirrettäville kovariaateille. Mikäli `axistext`-argumentille ei anneta arvoja, käytetään `legendtext`-argumentin arvoja. Mikäli kumpaakaan ei anneta, käytetään oletusarvoja.

Jos riskitiheyskuvio piirretään malliobjektia käyttäen, tekstit vaaka-akselille otetaan oletuksena malliobjektista (Liite D.3, rivit 211, 228, 233–240, 309–318, 323; Liite D.4, rivit 338, 355, 430–440, 445). Jos taas kuva piirretään ilman malliobjektia, oletusarvoina käytettävät nimet otetaan ensisijaisesti `x`-datamatriisin sarakkeilta ja toissijaisesti `xp`-datamatriisin sarakkeilta tai `coefs`-vektorista (Liite D.2, rivit 4, 57–68).

Oletuksena vaaka-akselille tulevat samat nimet kuin muuttujilla on aineistossa. Kun käytössä on `coxph`-luokan objekti, tällaisia nimiä ei saa malliobjektista suoraan, vaan nimet saadaan sellaisina kuin ne funktiokutsussa esiintyvät. Nimiä muokataan, siten, että saadaan esiin sama muoto, kuin kovariaatilla on aineistossa. Niistä nimistä, jotka poikkeavat aineistossa olevista, poistetaan kaikki merkit ensimmäisestä pilkusta eteenpäin sekä kaikki aukeavaan kaarisulkuun asti. Lopuksi poistetaan sulkevat kaarisulut. Näin jäljelle jää vain aineistossa esiintyvä muuttujan nimi, vaikka muuttujalle olisi tehty jokin monimutkainenkin muunnos. (Liite D.3, rivit 228, 233–240.)

Vaaka-akselin nimet tasataan oikeaan reunaan ja kovariaattien jakaumia kuvaavat arvot tulostetaan keskistetysti. Nimien oikean reunan sijaintia voi muuttaa `axistextposition`-argumenttia käyttäen. Oletusarvo on -0.1. Tekstien sijainnin muuttaminen on tarpeellista, mikäli kuva on leveä tai nimet pitkiä. Arvoa muuttaessa kannattaa ottaa huomioon, että vaaka-akselin arvo minimin kohdalla on 0 ja maksimin kohdalla 1. (Liite D.2, rivit 4, 167–168.)

Esimerkissä C.5 on muutettu vaaka-akselin tekstejä ja esimerkissä C.12 tekstejä on siirretty, jotta ne eivät mene päällekkäin muuttujien tunnuslukujen kanssa.

A.2.4 Selitelaatikon tekstit ja sijainti: `legendtext` ja `legendlocation`

Kuvaan piirretään selitelaatikko, josta nähdään, minkä kovariaattien suhteellisen riskitiheyden jakaumia viivat kuvaavat. Selitelaatikon tekstejä voi muut-

taa `legendtext`-argumentilla. Tekstit annetaan vain piirrettäville kovariaateille. Mikäli `axistext` on käytössä, mutta `legendtext` ei, käytetään ensimmäisen arvoja myös selitelaatikossa. Jos kumpikaan argumenteista ei ole käytössä, käytetään oletusarvoja.

Jos funktiokutsussa käytetään malliobjektia, otetaan selitelaatikon tekstit siitä (Liite D.3, rivit 210, 228, 309–318, 323; Liite D.4, rivit 337, 430–440, 445). Jos kuvaaja piirretään ilman malliobjektia, oletusarvoina käytettävät nimet otetaan `xp`-datamatriisin sarakkeilta tai `coefs`-vektorista (Liite D.2, rivit 3, 57–68). Oletuksena selitelaatikkoon tulee näkyviin myös kovariaatin muunnos.

Selitelaatikon paikan kuvassa voi määrätä käyttämällä `legendlocation`-argumenttia. Sille annetaan paikka siten kuin funktio `legend` sen vaatii: esimerkiksi "topleft", "topright" tai "bottom". Oletuksena on "top". Mikäli laatikkoa haluaa muokata, esimerkiksi pienentää tekstejä tai piirtää laatikon rajat näkyviin, selitelaatikon voi lisätä jälkikäteen. Tällöin argumentille `legendlocation` tulee antaa arvoksi `NULL` ja selitelaatikon voi lisätä itse käyttämällä kuvan piirtämisen jälkeen funktiota `legend`. (Liite D.2, rivit 4, 204–205.)

A.2.5 Referenssin osoittaminen: `reftick`, `refline`, `refline.col`, `refline.lwd`, `refline.lty`

Suhteellisen riskitiheyden (logaritmin) arvo vertailukohdassa on sen referenssi. Mikäli kuvataan suhteellinen riskitiheys, referenssi on 1. Suhteellisen riskitiheyden logaritmin tapauksessa referenssi on 0. Vertailukohdan löytämistä kuvasta nopeuttaa, jos referenssi on korostettu kuvaan. Jos argumentin `reftick` arvo on `TRUE`, referenssin asteikkomerkki lihavoidaan. Mikäli asteikkomerkkiä ei halua lihavoida, tulee argumentin arvoksi antaa `FALSE`. Oletuksena referenssin asteikkomerkki lihavoidaan. (Liite D.2, rivit 5, 103, 110, 151–152.)

Referenssin voi korostaa myös piirtämällä sen kohtaan suoran argumentilla `refline`. Tällöin arvo `TRUE` piirtää suoran. Oletuksena arvo on `FALSE`, ja suoraa ei piirretä. (Liite D.2, rivit 5, 103, 110, 154–155.)

Referenssisuoralle voi määrittää värin argumentilla `refline.col`, sekä paksuuden ja tyylin argumenteilla `refline.lwd` ja `refline.lty` (Liite D.2, rivit 5–6, 154–155). Oletuksena referenssisuora on musta katkoviiva. Referenssisuora on piirretty liitteen C.6 kuvaan 16.

A.2.6 Piirtoargumentit `col`, `lwd`, `lty`, `pch`, `bg`, `pt.lwd` ja `cex`

Jotta eri kovariaattien suhteellisen riskitiheyden jakauman kuvaajat erotetaan toisistaan, tulee viivojen tai vaihteluväliä ja kvartiileja kuvaavien pisteiden olla erilaisia. Viivat voidaan esimerkiksi piirtää eri värein. Myös eri viivatyylien käyttö voi olla hyödyllistä. Mikäli kuvia ei tulosteta väreissä, viivojen erottelu on helpointa erilaisia pisteitä ja tyylejä käyttämällä. Viivojen paksuuden ja pisteiden koon muuttaminen voi olla tarpeellista kuvan selkeyttämiseksi.

Viivan väri, paksuus ja tyyli määritetään käyttämällä vastaavassa järjestyksessä argumentteja `col`, `lwd` ja `lty`. Pisteen väri, muoto ja koko määrite-

tään käyttämällä argumentteja `col`, `pch` ja `cex`. Siis viiva ja piste ovat aina samanväriset.

Viivan paksuuden ja tyylin oletusarvo on 1 eli paksuntamaton yhtenäinen viiva. Myös argumentin `cex` oletusarvo on 1. (Liite D.2, rivit 8–9, 26, 46–48, 160–161.)

Väri voidaan R-ympäristössä määrittää esimerkiksi joko numerolla tai käyttämällä värin nimeä, kuten ”hotpink”. Oletuksena kaikki viivat piirretään eri värein käyttäen numeroita yhdestä eteenpäin. Tällöin ensimmäisenä piirrettävän kovariaatin väri on musta, joka on numero 1. Mikäli argumentille `col` annetaan arvoksi värit määrittävä vektori, sen arvot kierrätetään. (Liite D.2, rivit 8, 26, 54–55, 160–161.)

Kierrättäminen tarkoittaa, että annettua vektoria toistetaan niin kauan, että saavutetaan pituus, joka on piirrettävien kovariaattien määrä. Mikäli vektorista tulee liian pitkä, katkaistaan loppu pois. Jos esimerkiksi viivan värille annetaan arvoksi ”red”, jokaisesta viivasta piirretään punainen. Jos taas väriksi määritellään `c(`”red”, ”blue”, ”hotpink”`)`, ja piirrettäviä kovariaatteja on kaksi, piirretään punainen ja sininen viiva. Jos kovariaatteja on neljä, piirretään ensimmäinen ja viimeinen punaisella, toinen sinisellä ja kolmas pinkillä. Kaikissa muissakin tässä alaluvussa esitellyissä argumenteissa käytetään arvojen kierrättämistä. Kierrättäminen on toteutettu käyttäen `rep`-funktiota.

Pisteen merkki eli muoto voidaan määrittää käyttämällä numeroita tai antamalla merkki itse, esimerkiksi ”*”. Oletuksena kaikille kovariaateille piirretään erilaiset pisteet, jotka määritetään käyttäen numeroita nollasta eteenpäin. Ensimmäisenä piirrettävän kovariaatin merkki on neliön mallinen reunus. Mikäli piste on tyyliltään sellainen (numeroltaan 21–25), että sisus ja reuna voidaan värittää erikseen, määrää `col` reunan värin, `pt.lwd` reunan paksuuden ja `bg` sisuksen värin. Oletuksena pisteen reunus on paksuntamaton ja sisus läpinäkyvä. (Liite D.2, rivit 8–9, 26, 49–50, 52–53, 161–162.)

Viivojen värille ja pisteiden muodolle on järkevää antaa eri arvo jokaiselle kovariaatille, jolloin viivat on helpompi erottaa toisistaan. Viivojen paksuuden sekä pisteiden koon kohdalla voi taas saman arvon käyttäminen antaa selkeimmän lopputuloksen.

Luottamusvälit kovariaatin suhteellisen riskitiheyden jakaumalle piirretään käyttäen samoja värejä ja pistetyyppejä kuin suhteellisen riskitiheyden jakauman kuvaajalle on käytössä. Ainoastaan viivan tyyppi muuttuu: oletuksena jakauma piirretään yhtenäisellä viivalla ja luottamusvälit katkoviivalla. (Liite D.2, rivit 170–177.)

Selitelaatikon piirtämiseen käytettävä koodi on liitteen D.2 riveillä 204–205. Selitelaatikkoon piirretään viivat ja pisteet samanlaisina kuin kuvaan. Laatikon piirtämistä on käsitelty enemmän liitteessä A.2.4.

Liitteen D.2 riveillä 141–142 `col`, `lwd`, `lty`, `pch` argumentit ovat turhia, sillä `plot`-funktion tyyppinä on ”n”, eikä silloin piirretä annettua aineistoa. Funktiolla `plot` luodaan vaan pohja kuvalle. Kuvaajat piirretään käyttäen `lines`-funktiota rivillä 160.

A.2.7 Kuvan muokkaaminen muita argumentteja käyttäen: '...'

Riskitiheyskuvio piirretään `rankhazardplot`-funktion sisällä käyttäen `plot`-, `lines`- ja `points`-funktioita. Näillä funktioilla on sellaisiakin argumentteja, joita funktiolle `rankhazardplot` ei ole listattu. Niitä argumentteja pääsee hyödyntämään, kun funktion argumenttina on mukana kolme pistettä '...'.

Kolmen pisteen käyttö tekee funktioiden kirjoittamisesta yksinkertaista, sillä funktion kirjoittajan ei tarvitse ottaa huomioon kaikkia potentiaalisten käyttäjien mahdollisia toiveita. Kaikki hyödynnettävän funktion monipuoliset ominaisuudet tulevat käyttöön vähällä vaivalla.

Kolme pistettä esiintyy `rankhazardplot.default`-funktion sisällä `plot`-funktiossa rivillä 144, `lines` ja `points`-funktiossa riveillä 160–162 sekä vastaavasti luottamusvälejä piirtäessä riveillä 171–176. Mikäli funktiolle `rankhazardplot.coxph` tai `rankhazardplot.cph` annetaan ylimääräisiä argumentteja, ne välittävät ne vain eteenpäin `rankhazardplot.default`-funktiolle.

Hyödyllisin kolmeen pisteeseen sisältyvä argumentti on kuvalle otsikon lisäävä `main`-argumentti. Myös esimerkiksi piirrettävien viivojen päätepisteen muotoa voi muuttaa käyttämällä argumenttia `lend`.

Koska suhteellisen riskitiheyden jakauman kuvaajat piirretään silmukan sisällä, käytetään kolmeen pisteeseen sisältyvillä argumenteilla samaa arvoa jokaiselle kuvaajalle. Vaikka funktiolle antaisi siis vektorin argumentille `lend`, vain ensimmäistä arvoa käytetään kaikille kuvaajille.

A.3 Kuvaajan muokkaaminen `par`-funktiota käyttäen

Oletusasetuksilla piirrettäessä vaaka-akselille kuvaan mahtuu neljän kovariaatin tunnusluvut. Mikäli haluaa piirtää kuvaan enemmän suhteellisen riskitiheyden jakaumia, voi rivien määrää lisätä `par`-funktion `mar`-argumenttia käyttäen. Sen käyttäminen on hyödyllistä myös, jos piirtää vähemmän kuvaajia ja haluaa kaventaa marginaaleja.

Argumentille `mar` annetaan arvoksi vektori, johon tulee neljä alkiota: kuvan alle, vasemmalle puolelle, ylle ja oikealle puolelle jätettävien rivien määrä. Oletusarvo on `c(5.1, 4.1, 4.1, 2.1)`. Yksi kuvan alle jäävistä riveistä käytetään asteikkomerkkien piirtämiseen. Mikäli kovariaatteja kuvaavat nimet ovat pitkiä, tilaa voi lisätä suurentamalla kuvan vasemmalle puolelle jätettävää tilaa.

Kuvia voi myös piirtää useampia samaan kehykseen. Argumenteille `mfrow` ja `mfcol` annetaan vektori, joka määrittää kuinka monta riviä ja saraketta kuvia tulostetaan. Ensimmäisellä argumentilla kuvat tulostetaan riveittäin ja jälkimmäisellä sarakkeittain.

Liitteen C.4 kuvassa 14 on esitetty kaksi kuvaa päällekkäin samassa kehyksessä. Sivulta 81 alkavassa koodissa on esitetty, kuinka kuvat on aseteltu kuvaan `mfrow`-argumenttia käyttäen, ja kuinka kovariaattien tunnusluvut on mahdutettu kuvaan käyttämällä `mar`-argumenttia. Esimerkin C.7 kuvasta 17 voi huomata myös päinvastaisen käytön `mar`-argumentille: Koska kuvaan on piirretty saman kovariaatin eri muunnoksia, riittää esittää tunnusluvut alku-

peräisistä ja muunnetuista arvoista. Viimeiset tunnusluvut olisivat vastaavat kuin ensimmäiset, joten ne voi jättää kuvasta pois sallimalla vain kahden tunnuslukurivin mahtumisen kuvaan.

Funktiolla `par` on paljon muitakin argumentteja, joiden käyttöön voi tutustua sen dokumentaatiosta.

A.4 Suhteellisen riskitiheyden jakauman piirtämiseen käytettävien arvojen määrittäminen

Seuraavaksi käydään läpi, kuinka suhteellisen riskitiheyden jakauman piirtämiseen tarvittavat arvot lasketaan funktiossa `rankhazardplot`. Teksti on tarkoitettu selventämään funktion koodia, ja liitteessä olevan koodin seuraaminen on oleellista kokonaiskäsityksen saamiseksi.

Kuvaajan piirtämiseksi tarvitaan kaksi koordinaattia kullekin pisteelle. Suhteellisen riskitiheyden jakauman tapauksessa vaaka-akselilla oleva koordinaatti on $[0, 1]$-välille skaalattu järjestysluku, joka määritetään kovariaatin arvojen perusteella. Kovariaatin jakaumaa kuvataan vaihteluvälin ja kvartiilien avulla, ja nämä tunnusluvut merkitään vaaka-akselille suuruusjärjestyksessä.

Pystyakselille lasketaan suhteellinen riskitiheys, jota varten tarvitaan myös vertailukohta. Mikäli vertailukohtaa ei anneta, se määritetään `rankhazardplot`-funktion sisällä. Oletuksena käytetään mediaanin arvoa. Suhteelliselle riskitiheydelle lasketaan myös luottamusvälit, mikäli muuttuja on alkuperäinen eli aineistossa oleva, yksinkertainen muunnos tai faktori.

Faktorit aiheuttavat monimutkaisuutta kvartiilien ja mediaanien määrittämiseen, sillä niiden tasoilla ei välttämättä ole järjestystä. Tällöin kyseiset tunnusluvut eivät ole määritelty. Kuvaajan kannalta tämä ei haittaa, ja R-kieli käyttää tällöin faktorin tasojen järjestystä, mikä on usein aakkosjärjestys.

A.4.1 Skaalattujen järjestyslukujen määrittäminen

Skaalatut järjestysluvut toimivat vaaka-akselin koordinaatteina. Ne määritetään piirrettävälle kovariaatille siten, että ensin selvitetään, kuinka monta havaintoa kovariaatista on saatu. Sitten `seq`-funktiota käyttäen luodaan sen pituinen vektori välille $[0, 1]$. Tällöin arvot ovat tasavälisiä. Tämä toistetaan jokaiselle piirrettävälle kovariaatille for-silmukassa, jossa myös kuvan piirtäminen tapahtuu. (Liite D.2, rivi 133, 136–137, 160, 171 ja 174.)

A.4.2 Vaihteluvälin ja kvartiilien määrittäminen

Kuvassa kaikki kovariaatit on siirretty $[0, 1]$-välille, eikä vaaka-akselilta voi ilman vaihteluväliä ja kvartiileja nähdä, millaisia arvoja kustakin muuttujasta on havaittu. Näiden tunnuslukujen määrittäminen on toteutettu siten, että sama tyyli sopii sekä tavallisille muuttujille että faktoreille.

Määrittäminen tehdään jokaiselle kovariaatille erikseen, ja se tapahtuu kolmessa vaiheessa. Tavoitteena on siis löytää indeksit, joista halutut havainnot

löytyvät aineistosta. Ensin selvitetään, mikä on kovariaatin pienimmän arvon indeksi, mikä toiseksi pienimmän, kolmanneksi pienimmän jne. Tämän tekee funktio `order`. Puuttuvat havainnot saavat suurimmat järjestysluvut. Tällä tavoin ne on helppo jättää käyttämättä piirtovaiheessa. (Liite D.2, rivit 124–125.)

Toiseksi määritetään, mitkä ovat kovariaatin minimin, maksimin sekä kvartiilien järjestysluvut aineistossa. Ne riippuvat ainoastaan havaintojen määrästä, ja puuttuvat havainnot on jätetty laskuista pois. Minimi on aina 1, ja maksimi havaintojen lukumäärä. Mikäli havaintoja on parillinen määrä, järjestysluku on kokonaisluku. Mikäli havaintoja on pariton määrä, käytetään järjestyslukujen kvartiilien arvot pyöristettynä alaspäin kokonaislukuun. (Liite D.2, rivit 138, 164–165.)

Ensimmäisessä ja toisessa vaiheessa saatujen arvojen avulla voidaan selvittää, missä indeksissä kovariaatin minimi, alakvartiili, mediaani, yläkvartiili ja maksimi sijaitsevat. Niissä indekseissä olevat kovariaatin arvot valitaan ja tulostetaan vaaka-akselille. Mikäli arvot ovat numeerisia, ne pyöristetään kolmen merkitsevän numeron tarkkuudelle, jotta ne eivät mene päällekkäin. (Liite D.2, rivit 164–168.)

A.4.3 Vertailukohdan oletusarvo ja sen vaihtaminen

Vertailukohtien määrittäminen tapahtuu tiivistetysti siten, että numeerisille muuttujille käytetään mediaania ja faktoreille vertailutasoa. Faktoreiden käyttö on mahdollista, mikäli `rankhazardplot`-funktiota kutsutaan malliobjektia käyttäen. Vertailukohtien määrittäminen toimii samalla tavalla käytettävän malliobjektin luokasta riippumatta.

Malliobjektia käytettäessä vertailuarvot lasketaan antamalla vertailukohdat `predict`-funktiolle. Se asettaa joitakin rajoitteita toteutukseen: `predict` vaatii aineiston datamatriisina, jossa on kaikki mallissa olevat muuttujat, joten tällainen datamatriisi luodaan funktiossa ja täytetään halutuilla arvoilla. Datamatriisiin sijoitetaan numeerisille muuttujille mediaani ja faktoreille vertailutaso. Tämän jälkeen tarkistetaan, onko piirrettäville kovariaateille annettu vertailukohdat argumentissa `refpoints`. Jos on, selvitetään, minkä piirrettävien kovariaattien vertailukohdat halutaan vaihtaa. Argumentti `select` sisältää piirrettävien kovariaattien järjestysnumerot. (Liite D.3, rivit 228, 245–262; liite D.4, rivit 355, 365–383)

Jos kaikki vaihdettavat vertailuarvot ovat numeerisia, on `refpoints` numeerinen vektori. Jos yksikin on merkkijono eli faktorin taso, ovat kaikki annetut vertailukohdat merkkejä. Vektorissa `refpoints` olevat arvot muutetaan samaan muotoon kuin ne aineistossa ovat. Tällöin `predict`-funktio osaa käsitellä arvoja oikein. (Liite D.3, rivit 263–270; liite D.4, rivit 384–391)

Vertailukohtien osittainen vaihtaminen ei toimi tässä versiossa täydellisesti: jos vaihdettavissa vertailukohdissa on mukana faktori, voi funktio kaatua sopimattoman kokoisiin sijoituksiin tai vertailukohtien sekoittumiseen muuttujien välillä. Virhe tapahtuu koodissa D.3 riveillä 266–269 ja koodissa D.4

riveillä 387–390. Virhe korjaantuu vaihtamalla kaikki noilla riveillä olevat koodin osat `is.element(select[change], ...)` muotoon `is.element(select, ...)`. Tällöin kyseisellä kutsulla syntyvä, loogisilla muuttujilla valinnan tekevä, vektori on aina yhtä pitkä kuin vektori `refpoints`. Siten sijoitettavasta muuttujasta tulee oikean kokoinen, ja luvut valitaan oikeasta kohdasta.

Vertailukohtien oletusarvojen määrittäminen tapahtuu hieman eri tavalla, mikäli `rankhazardplot`-funktiota kutsutaan käyttäen `coefs`-argumenttia. Tällöin kaikki piirrettävät muuttujat ovat numeerisia, ja vertailukohta voidaan määrittää jokaiselle kovariaatille mediaanina. Mikäli käyttäjä antaa vertailukohdat, ne tulee antaa jokaiselle piirrettävälle muuttujalle. Vertailuarvot lasketaan kertoimien ja vertailukohtien tulona. (Liite D.2, rivit 35–41.)

Jos käytössä on `xp-` tai `confinterval`-argumentti, vertailukohtien sijaan käytetään vertailuarvoja, jotka sisältyvät `confinterval`-argumenttiin automaattisesti.

Liitteessä A.4.2 käsitellyt kvartiilit määritetään aitoina havaintoina. Mikäli havaintoja on parillinen määrä, mediaani on jokin havaituista arvoista vain, jos kaksi keskimmäistä arvoa ovat samat. Tällöin vertailuarvon laskemisessa ei käytetä vaaka-akselille tulostettavaa mediaania. Siksi suhteellisen riskitiheyden referenssi ei ole aina kuvan mediaanissa, vaikka suhteellisen riskitiheyden jakauma piirrettäisiin vertailukohtien oletusarvoja käyttäen.

A.4.4 Suhteellisen riskitiheyden laskeminen

Liitteestä A.4.3 käy ilmi, että vertailuarvot voidaan laskea kahdella eri tavalla: funktiota `predict`-käyttäen tai kertoimien ja vertailukohtien tulona. Tulos ei kuitenkaan ole sama, sillä `predict`-funktio palauttaa arvot, joista on kovariaateittain vähennetty mallin sovituksessa käytetyn aineiston termittäisten ennusteiden keskiarvot. Se ei haittaa, sillä kun termittäiset ennusteet lasketaan samalla tavalla havainnoille ja vertailukohdille, keskistystermit kumoutuvat laskussa. Olkoon kovariaatin j keskistystermi c_j. Tällöin suhteellinen riskitiheys `predict`-funktion antamien arvojen perusteella on

$$\begin{aligned}\exp\Big(\beta_j\big(x_{i,j} - c_j - (x_{\mathrm{ref},j} - c_j)\big)\Big) &= \exp\Big(\beta_j\big(x_{i,j} - c_j - x_{\mathrm{ref},j} + c_j\big)\Big) \\ &= \exp\big(\beta_j(x_{i,j} - x_{\mathrm{ref},j})\big),\end{aligned}$$

mikä on sama kuin kaavassa (4) esitetty muoto suhteelliselle riskitiheydelle.

Funktiot `rankhazardplot.coxph` ja `rankhazardplot.cph` laskevat ennusteet keskistettyinä ja `rankhazardplot.default` tulona. Syy tähän kahtiajakoon on se, että oletusfunktiolle ei anneta mallia, jota tarvitaan `predict`-funktion kutsumiseen, joten ennusteet tulee laskea yksinkertaisemmin kertolaskuna. Toisaalta taas `predict`-funktio on monipuolisempi kuin kertolasku, sillä sitä käyttäen saadaan ennusteet monimutkaisillekin muunnoksille, kuten splineille.

Termittäisten ennusteiden laskeminen molemmilla malliobjekteilla toimii taas pääpiirteissään samoin. Aineisto `x` sisältää vain mallissa olevien muuttu-

jien arvot, ja `refs` sisältää vertailukohdat, kuten luvussa A.4.3 on kuvattu. Ennusteet lasketaan aina kaikille mallissa oleville muuttujille, vaikka kaikkia ei piirrettäisikään. Jatkokäsittelyä varten vertailukohtien termittäiset ennusteet muutetaan vektoriksi ja havaintojen termittäiset ennusteet datamatriisiksi. Lasketuista arvoista vain piirrettävien kovariaattien arvoja käytetään kutsuttaessa oletusfunktiota, sillä `rankhazardplot.default` piirtää kaikkien `xp`-argumentissa saatujen kovariaattien suhteellisen riskitiheyden jakaumat. (Liite D.3, rivit 273–281, 321–322; liite D.4, rivit 394–400, 443–444.)

Virheilmoitus `rankhazardplot.default`-funktiossa varmistaa, että vertailuarvot on annettu, mikäli funktiota kutsutaan `xp`-argumenttia käyttäen. Ainoastaan `coefs`-argumentilla kutsuttaessa sekä termittäiset ennusteet että vertailuarvot lasketaan kertolaskulla. Matriisien kertolaskua ja erotusta käyttäen termittäisten ennusteiden avulla lasketaan ensin suhteellisen riskitiheyden logaritmi, joka muutetaan suhteelliseksi riskitiheydeksi, mikäli funktiota kutsuttaessa kuvion tyypiksi on määritetty ”hazard”. (Liite D.2, rivit 32–33, 43–44, 70–71, 79–81.)

A.4.5 Suhteellisen riskitiheyden luottamusvälin laskeminen

Luottamusvälien laskeminen on toteutettu kertolaskulla, sillä tällöin laskussa käytettävät Coxin mallin kertoimet on helppo muuttaa luottamusvälinsä ala- ja ylärajan mukaisiksi, ja laskut voidaan tehdä niitä käyttäen. Tällä hetkellä `rankhazardplot`-funktiolla ei voi kuitenkaan laskea luottamusvälejä kaikille muunnoksille, kuten splineille. Laskeminen onnistuu alkuperäisille kovariaateille, niiden yksinkertaisille muunnoksille ja faktoreille. Luottamusvälien laskemista varten on tehty funktiot `coxph_CI` (Liite D.5) ja `cph_CI` (Liite D.6), joita vastaavasti kutsutaan funktioista `rankhazardplot.coxph`- ja `rankhazardplot.cph`.

Funktioiden `coxph_CI` ja `cph_CI` argumentit ovat malliobjekti, aineisto, Coxin mallin kertoimet sekä vertailukohdat. Ainoa ero funktioiden toiminnassa ovat koodirivit, joissa hyödynnetään malliobjektia.

Funktioille annettava aineisto poikkeaa muissa `rankhazard`-paketin funktioissa käytettävistä aineistoista. Se on samanlainen, jonka funktiot `coxph` ja `cph` sisältävät, mikäli argumentti `x = TRUE`. Aineisto annetaan `rankhazardplot`-funktiolle argumenttina `x_CI` tai malliobjektissa. Jokaisesta kovariaatista tulee aineistoon yhtä monta saraketta kuin kertoimia on estimoitu. Esimerkiksi faktorit esitetään indikaattorimuuttujina. Yksinkertaisten muunnosten arvot ovat aineistossa muunnettuina, eivät alkuperäisinä. Luottamusvälit lasketaan faktoreille sekä kaikille muuttujille, joiden arvot ovat yhtenä sarakkeena. (Liite D.5, rivit 466–467, 471–478; liite D.6, rivit 532–533, 537–544.)

Termittäiset ennusteet lasketaan kertoimien ja aineiston arvojen tulona. Tämän jälkeen faktoreiden eri tasojen ennusteet yhdistetään ja kovariaatin nimeksi muutetaan mallin kutsussa käytetty muoto. Faktorien termittäiset ennusteet sisältävät jokaiselle vertailutasosta poikkeavalle tasolle muuttujan, joka saa tason estimoidun kertoimen arvon, jos havaintoyksikön arvo on kyseinen

taso, muuten arvon nolla. Kun nämä muuttujat lasketaan riveittäin yhteen, saa kukin havaintoyksikkö faktorin tasoa vastaavan kertoimen arvon. Jos arvo on nolla, kyseessä on vertailutaso. Jos faktorilla on vain kaksi tasoa, vertailutasosta poikkeavan tason ennusteet kopioidaan faktorin nimellä otsikoituun muuttujaan. (Liite D.5, rivit 469, 480, 484-488; liite D.6, rivit 535, 546, 550–554.)

Lisäksi faktorista kootaan alkuperäinen muuttuja, jossa tasot ovat merkkijonoja. Tämä onnistuu käyttäen indikaattorimuuttujia ja mallista saatuja faktorien tasoja. Tämän jälkeen aineistossa olevat merkkijonoja sisältävät muuttujat pakotetaan faktoreiksi, joilla on sama vertailutaso kuin mallin sovituksessa käytetyllä muuttujalla. (Liite D.5, rivit 481–484, 490–505; liite D.6, rivit 547–550, 556–571.)

Sitten termittäisistä ennusteista ja aineistosta valitaan vain ne muuttujat, joille luottamusvälit on laskettu. Lopuksi lasketaan vertailuarvot. Jotta suhteellinen riskitiheys lasketaan faktoreille oikein, tulee vertailuarvon olla sama kuin vertailutasolla, eli nolla. Tämä varmistetaan funktion lopussa turhaan, sillä faktoreiden vertailukohdat on annettu nolliksi jo funktiota kutsuttaessa. Funktio palauttaa aineiston ja termittäiset ennusteet, joilla on nyt vain yksi sarake muuttujaa kohden. Lisäksi palautettavassa listassa on vertailuarvot, sekä tieto siitä, mille alkuperäisen aineiston muuttujille luottamusvälit on laskettu. (Liite D.5, rivit 510–521; liite D.6, rivit 576–587.)

Ennen funktion `coxph_CI` tai `cph_CI` kutsumista selvitetään Coxin mallin kertoimien luottamusvälit, ja vertailukohdat sisältävä datamatriisi muunnetaan matriisiksi. Tätä ennen faktoreiden vertailukohdiksi annetaan nolla. Sitten kolmea eri kutsua käyttäen lasketaan termittäiset ennusteet ja niiden luottamusvälien ala- sekä ylärajat. Ennusteet lasketaan toiseen kertaan (ensimmäiset on laskettu `predict`-funktiolla), jotta jakauman ja luottamusvälien piirtämisessä käytettävä aineisto on sama. Oletuksena käytettävä aineisto sisältää vain mallin sovituksessa mukana olevien havaintoyksikköjen arvot. (Liite D.3, rivit 225–226, 283–293; liite D.4, rivit 352–353, 402–412.)

Tuloksista kootaan lista `confinterval`, joka sisältää aineiston, termittäiset ennusteet sekä niiden luottamusvälien ala- ja ylärajat ja samat vertailuarvoille. Luottamusvälien piirtämiseen tarvittavat arvot lasketaan aina, ja ne palautetaan, mikäli argumentin `return` arvo on `TRUE`. (Liite D.3, rivit 295–296, 332–333; liite D.4, rivit 415–416, 454–455.)

On mahdollista, että kun funktiota `rankhazardplot` kutsutaan malliobjektia käyttäen, luottamusvälejä pyydetään muuttujalle, jolle niitä ei voida laskea. Tällöin funktion suoritus keskeytetään ja tulostetaan virheilmoitus ”Confidence intervals cannot be calculated for selected covariates”. Luottamusvälit piirretään, jos `confint = TRUE` ja `draw = TRUE`. Tällöin riskitiheyskuvion piirtämisen tekevää `rankhazardplot.default`-funktiota kutsutaan saatua `confinterval`-objektia käyttäen. (Liite D.3, rivit 297–308, 320–324; liite D.4, rivit 417–428, 442–446.)

Suhteelliset riskitiheydet luottamusväleille lasketaan `rankhazardplot.de-`

`fault`-funktiossa argumentin `confinterval` sisältävien tietojen perusteella. Ensin tiedot kerätään listasta erillisiksi objekteiksi. Sen jälkeen toimitaan samalla tavalla kuin suhteellisen riskitiheyden laskemisen kanssa: ensin lasketaan suhteellisen riskitiheyden logaritmi luottamusväleille, ja se muunnetaan suhteelliseksi riskitiheydeksi, mikäli näin halutaan. (Liite D.2, rivit 11–19, 73–78, 82–85.)

A.5 Riskitiheyskuvion piirtäminen

Riskitiheyskuvion piirtämiseen tarvitaan useita tietoja, joita on käsitelty liitteessä A.4. Ennen kuvaajien piirtämistä on selvitetty jokaisen muuttujan arvojen suuruusjärjestyksen indeksit. Ne ovat sarakkeilla matriisissa `ind`. (Liite D.2, rivit 122–125.)

Riskitiheyskuvion piirtäminen tapahtuu for-silmukan sisällä. Ensimmäinen piirrettävä asia on kuvan pohja: kuvalle määritetään vaaka- ja pystyakselin vaihteluvälit, pystyakselin otsikko, mahdollinen kuvan otsikko sekä se, onko pystyakselin asteikko logaritminen vai ei. Edes akseleita ei piirretä kuvaan, vaan ne lisätään seuraavaksi. (Liite D.2, rivit 140–144.)

Asteikkomerkit piirretään vaaka-akselille kohtiin 0, 0.25, 0.5, 0.75 ja 1, mutta otsikot jätetään tässä vaiheessa pois. Pystyakselin asteikkomerkit ja selitteet tehdään myös kahdella eri komennolla. Näin voidaan itse valita, mille asteikkomerkeille selite lisätään. Muuten selitteitä lisätään jokaiselle merkille, jolle selite alhaalta päin aloittaen mahtuu tulostumaan riittävällä välillä edelliseen. Mikäli käyttäjä yrittää tulostaa liian lähekkäin olevia selitteitä merkeille, niistä osa jää pois. Kuvan rajat viimeistellään funktiolla `box`. (Liite D.2, rivit 146–149.)

Mikäli referenssi halutaan korostaa kuvaan, voidaan joko referenssin asteikkomerkki lihavoida tai lisätä kuvan pohjalle referenssisuora. Tämän jälkeen kuvan pohja on valmis, ja varsinainen kuvaajien piirtäminen voidaan aloittaa. (Liite D.2, rivit 151–156.)

Jakauman kuvaaja piirretään `lines`-funktiolla, jolle vaaka-akselin koordinaatit ovat skaalatut järjestysluvut ja pystyakselille valitaan suhteellisen riskitiheyden arvot kovariaatin arvojen suuruusjärjestyksessä matriisia `ind` käyttäen. Kovariaatin vaihteluvälin ja kvartiilien kohtaan piirretään pisteet, jotka ovat vaaka-akselin asteikkomerkkien kohdalla, ja suhteellinen riskitiheys on laskettu vaaka-akselille merkittävälle arvolle. (Liite D.2, rivit 160–162.)

Liitteessä A.4.2 selvitetyt kovariaatin tunnusluvut tulostetaan vaaka-akselille. Kovariaatin nimi tasataan oikeaan reunaan, ja tunnuslukujen arvot tulostetaan keskistetysti. Tämän jälkeen siirrytään seuraavaan kovariaattiin, ellei tarkoituksena ole piirtää luottamusvälejä. (Liite D.2, rivit 164–168.)

Luottamusvälit piirretään samalla periaatteella kuin suhteellisen riskitiheyden jakauma. Suhteellisten riskitiheyksien sijaan piirretään luottamusvälin ala- ja ylärajojen jakaumat kahdella eri kutsulla. Luottamusvälit piirretään erilaisella viivalla kuin suhteellisen riskitiheyden jakauma. Jos jakauman piirtämi-

seen on esimerkiksi käytetty yhtenäistä viivaa, luottamusvälit piirretään katkoviivalla. (Liite D.2, rivit 170–178.)

A.6 Funktion `rankhazardplot` palauttamat arvot

Uudistettuun `rankhazardplot`-funktioon on tehty ominaisuus, jonka avulla eri malleissa olevien kovariaattien suhteellisen riskitiheyden jakaumia voi piirtää samaan kuvaan. Versiossa 0.8-1 se onnistui argumenttia `coefs` käyttäen numeerisille muuttujille, joille estimoidaan vain yksi kerroin. Periaatteessa olisi ollut mahdollista piirtää yli kaksitasoisien faktoreiden suhteellisen riskitiheyden jakaumia. Se olisi vaatinut, että käyttäjä olisi laskenut itse termittäiset ennusteet ja vertailuarvot ennen `rankhazardplot`-funktion kutsumista `xp`-argumenttia käyttäen.

Nyt tämä vaihe on automatisoitu: kuvaajien yhdistäminen eri malleista on toteutettu palautettavien arvojen avulla. Funktiot `rankhazardplot.coxph` ja `rankhazardplot.cph` palauttavat suhteellisen riskitiheyden jakauman piirtämiseen tarvittavat arvot, eli kovariaattien arvot, termittäiset ennusteet ja vertailuarvot. Lisäksi palautetaan lista, joka sisältää luottamusvälien piirtämiseen tarvittavat arvot. Funktion palauttamat arvot on esitelty dokumentaatiossa, joka on liitteessä B.5. (Liite D.3 rivit 295–296, 333; liite D.4, rivit 415–416, 455)

Dokumentaatiossa on liitteessä C.7 esimerkki, jossa piirtämiseen on käytetty funktion palauttamia arvoja. Kuvaan 17 on yhdistetty yhden kovariaatin muunnokset kolmesta eri mallista samaan kuvaan. Palautetuista arvoista täytyy valita halutut ja koota niistä `rankhazardplot.default`-funktiolle annettavat arvot. Tämä vaatii hieman työtä, mutta kaikki jakaumat mahtuvat automaattisesti kuvaan.

A.6.1 Arvojen palauttamisen ja kuvan piirtämisen kontrollointi: `return` ja `draw`

Mikäli suhteellisen riskitiheyden jakaumien kuvaajia haluaa yhdistää eri malleista samaan riskitiheyskuvioon, ja piirrettävien kovariaattien joukossa on esimerkiksi faktori, tulee kuva piirtää käyttäen `xp`-argumenttia. Mikäli kaikki muuttujat ovat lineaarisia, voi käyttää myös `coefs`-argumenttia. Silloin piirtäminen on yksinkertaisempaa kuin mitä seuraavassa esitetään.

Ensin termittäiset ennusteet tulee laskea kutsumalla `rankhazardplot`-funktiota malliobjektia käyttäen. Termittäiset ennusteet saadaan jatkokäsittelyä varten asettamalla argumentin `return` arvoksi `TRUE`. Funktio palauttaa liitteessä A.6 esitettävät arvot. Piirtokutsua varten tulee näistä palautetuista arvoista valita kuvaan halutut. (Liite D.3, rivit 215, 332; liite D.4, rivit 342, 454.)

Yhtä kuvaa varten tulee tällöin useita funktiokutsuja. Oletusarvon mukaan riskitiheyskuvio piirretään joka kerta. Kuvan piirtämisen voi estää asettamalla argumentin `draw` arvoksi `FALSE`. Tällöin kuvan piirtämisen suorittavaa

`rankhazardplot.default`-funktiota ei kutsuta ollenkaan. (Liite D.3, rivit 214, 320; liite D.4, rivit 341, 442.)

Jos sekä `return`- että `draw`-argumentin arvo on `FALSE`, `rankhazardplot`-funktion kutsu ei palauta tai tulosta mitään.

Esimerkki kuvaajien yhdistämisestä samaan riskitiheyskuvioon on esitetty liitteessä C.7 kuvassa 17. Koodi alkaa sivulta 87. Kuvassa on vertailtu eri muunnosten suhteellisen riskitiheyden jakaumia bilirubiini-kovariaatille. Kuvaa käsitellään enemmän luvussa 5.2.

A.7 Funktion `rankhazardplot` tuloste

Funktio `rankhazardplot.default` tulostaa suhteellista riskitiheyttä kuvaavia arvoja. Ensimmäisenä tulostetaan suhteellisen riskitiheyden vaihteluväli, mikä on hyödyllinen, kun halutaan piirtää useita keskenään vertailukelpoisia kuvia.

Toiseksi tulostetaan kunkin kovariaatin suhteellisen riskitiheyden vaihteluväli ja kvartiilit suuruusjärjestyksessä. Mikäli suhteellisen riskitiheyden kuvaaja on kasvava, kyseiset luvut ovat likimain kovariaatin suhteellisen riskitiheyden jakaumalle piirrettyjen pisteiden pystyakselin koordinaatit. Likimain siksi, että tulostettavat arvot ovat havaittuja suhteellisia riskitiheyksiä vain, kun havaintoja on pariton määrä, mutta pisteiden koordinaatit ovat aina peräisin aidoista havainnoista. (Liite D.2, rivit 117–132.)

Mikäli suhteellisen riskitiheyden jakauman kuvaaja on laskeva, ovat tulostetut luvut päinvastaisessa järjestyksessä. Tällöin pienimmällä kovariaatin arvolla saavutetaan suhteellisen riskitiheyden maksimi. On myös mahdollista, ettei suhteellisen riskitiheyden jakauma ole monotoninen, esimerkiksi jos kyseessä on monitasoinen faktori. Tällöin kaikki tulostetut arvot eivät ole välttämättä yhdistettävissä kuvaan piirrettyjen pisteiden koordinaatteihin.

Dokumentaation esimerkistä C.4 löytyy sivun 81 alareunasta tuloste kuvasta, jossa on sekä laskeva että kasvava kuvaaja. Kuva on sivulla 82.

Jos luottamusvälit piirretään, niille tulostetaan samanlainen taulukko kuin suhteellisille riskitiheyksille. Tulosteista on esimerkki sivulla 95 esimerkissä C.13.

B Julkaistu englanninkielinen rankhazardplot-dokumentaatio

B.1 Description

Creates a rank-hazard plot. Plots the relative hazards (or the logarithm of the relative hazards) for each covariate of a Cox proportional hazards model fitted by `coxph` or `cph`.

B.2 Usage

```
rankhazardplot(...)

## S3 method for class 'coxph'
rankhazardplot(coxphobj, data, select = NULL, refpoints = NULL,
    CI_level = 0.95, x_CI = NULL, confint = FALSE,
    legendtext = NULL, axistext = NULL, legendlocation = "top",
    axistextposition = -0.1, reftick = TRUE, refline = FALSE,
    refline.col = 1, refline.lwd = 1, refline.lty = 2,
    ylab = NULL, ylim = NULL, yticks = NULL, yvalues = NULL,
    plottype = "hazard", na.rm = TRUE, draw = TRUE,
    return = FALSE, col = NULL, lwd = 1, lty = 1, pch = NULL,
    cex = 1, bg = "transparent", pt.lwd = 1, ...)

## S3 method for class 'cph'
rankhazardplot(cphobj,  data, select = NULL, refpoints = NULL,
    CI_level = 0.95, x_CI = NULL, confint = FALSE,
    legendtext = NULL, axistext = NULL, legendlocation = "top",
    axistextposition = -0.1, reftick = TRUE, refline = FALSE,
    refline.col = 1, refline.lwd = 1, refline.lty = 2,
    ylab = NULL, ylim = NULL, yticks = NULL, yvalues = NULL,
    plottype = "hazard", na.rm = TRUE, draw = TRUE,
    return = FALSE, col = NULL, lwd = 1, lty = 1, pch = NULL,
    cex = 1, bg = "transparent", pt.lwd = 1, ...)

## S3 method for class 'default'
rankhazardplot(x, coefs = NULL, xp = NULL, refvalues = NULL,
    refpoints = NULL, confinterval = NULL, select = 1,
    legendtext = NULL, axistext = NULL, legendlocation = "top",
    axistextposition = -0.1, reftick = TRUE, refline = FALSE,
    refline.col = 1, refline.lwd = 1, refline.lty = 2,
    ylab = NULL, ylim = NULL, yticks = NULL,
    yvalues = NULL, plottype = "hazard", na.rm = TRUE,
    col = NULL, lwd = 1, lty = 1, pch = NULL,
```

```
cex = 1, bg = "transparent", pt.lwd = 1, ...)
```

B.3 Arguments

coxphobj

An object of class `coxph` created by function `coxph` from the package `survival`. The object should have been created with `x = TRUE`. Otherwise `x_CI` must be given.

cphobj

An object of class `cph` created by function `cph` from the package `rms`. The object should have been created with the option `x = TRUE`. Otherwise `x_CI` must be given.

data

A data frame that contains the covariates in the model. It can be the data used in fitting the Cox proportional hazards model or new data. It can contain more covariates than what there are in the model. Used with the argument `coxphobj` or `cphobj`.

select

A vector with the order numbers of the covariates to be plotted. The order numbers are defined by the order of the covariates in the model. It can be used with the argument `coxphobj`, `cphobj` or `confinterval`. When used with `coxphobj` or `cphobj` the default is `NULL` and all covariates in the model are plotted. With `confinterval` the default is the first covariate.

x

A data frame that contains the covariates to be plotted with one column for each covariate. Used with the argument `xp` or `coefs`. When used with `xp`, `x` defines the values to be printed on the x-axis. When used with `coefs`, `x` the predictions are computed as as a product of `coefs` and `x`. The dimensions of `x` and `xp` must be the same. The number of covariates given in `x` and the length of the `coefs` must be the same. If `return = TRUE` `x` is returned in the list. See section `Value`.

xp

A data frame that contains the predictions (”terms”) for the covariates to be plotted. The order of the covariates must be the same as in `x`. If `return = TRUE` `xp` is returned in the list. See section `Value`.

coefs

A vector of Cox regression coefficients for the covariates to be plotted. The order of the covariates must be the same as in `x`.

refpoints

A vector of reference points given in the same order as the covariates in the model. A reference point is the value of the covariate where the reference hazard is calculated to compute the relative hazards as a quotient of the hazard and the reference hazard. Consequently, at the reference point the value of the relative hazard is 1 (and the value of the logarithm of the relative hazard is 0). If the `select` argument is in use the reference points are given for the selected covariates only and in the same order as the selection made. If `NULL`, the medians of each covariate are used as reference points. With factors the default is the reference level of the factor. If the reference point for the selected covariate is `NA`, the default is used. When plotting the confidence intervals the reference point for factors can be changed only by re-leveling the factor.

refvalues

A vector of reference values. A reference value of a covariate is the predicted value ("terms") at the reference point (see `refpoints`). Used and needed only with the argument `xp`. Otherwise calculated by the `rankhazardplot` function with the values of the `refpoints`. If `return = TRUE` `refvalues` is returned in the list. See section `Value`.

CI_level

A number between 0 and 1 that defines the level of the confidence interval for (the logarithm of) the relative hazard. By default 0.95.

x_CI

A data frame of the covariate data. Needed if the `coxphobj$x` or `cphobj$x` does not exist. The number and the order of the columns must be same as `as.data.frame(coxphobj$x)` or `as.data.frame(cphobj$x)`.

confint

If `TRUE` the confidence intervals are plotted. Can be used with `coxphobj` or `cphobj`. By default `FALSE`.

confinterval

A list that contains all information to plot confidence intervals. If `return = TRUE` `confinterval` is returned in the list. See section `Value`.

legendtext

A vector of covariate names for the legend box. If `NULL` and `axistext = NULL`, the names are from the columns of `xp`, `coefs`, `attr(coxphobj$terms, "term.labels")` or `attr(cphobj$terms, "term.labels")`. If `NULL`, `axistext` is used if it is given. If the `select` argument is in use the names are given for the selected covariates only and in the same order as the selection made.

axistext

A vector of covariate names and units for the x-axis. If `NULL` and `legendtext = NULL`, the names are from the columns of `x` or the same as in the `data`. If `NULL`, `legendtext` is used if it is given. If the `select` argument is in use the names are given for the selected covariates only and in the same order as the selection made.

legendlocation

A keyword that determines the location where the legend box is printed. By default "top". See `Details` in the documentation for `legend`. If `NULL`, the legend box is not printed and it can be added by function `legend`.

axistextposition

A number that defines the x-coordinate where the axis texts are placed. Adjustment is right alignment. By default -0.1. The bigger the size of the plot is, the closer to zero the value can be. The maximum is 0.

reftick

By default `TRUE` and the reference tick is emboldened. For the relative hazard the tick is at 1 and for the logarithm of the relative hazard at 0.

refline

If `TRUE` the reference line is drawn. The line is horizontal at the same place as the reference tick (see `reftick`). By default `FALSE`.

refline.col

Defines the colour of the reference line, if `refline = TRUE`. By default 1. See documentation for `par`.

refline.lwd

Defines the width of the reference line, if `refline = TRUE`. By default 1. See documentation for `par`.

refline.lty

Defines the type of the reference line, if `refline = TRUE`. By default 2. See documentation for `par`.

ylab

A string that defines the label of the y-axis. When `plottype = "hazard"` the default is "relative hazard". When `plottype = "loghazard"` the default is the "logarithm of the relative hazard". See documentation for `plot`.

ylim

A vector that defines the range of the y-axis. By default the range is the minimum and the maximum of (the logarithm of) the relative hazards. See documentation for `plot.window`.

yticks

A vector that determines the places for the ticks that are plotted to the y-axis. When `plottype = "hazard"` it is recommended that evenly calculated ticks are used before and after the reference e.g. `c(seq(0.5, 1, by=0.1),2:7)`, due to the logarithmic scale of the y-axis. If `NULL` the ticks are computed using `pretty`.

yvalues

A vector that determines which values are printed on the y-axis. If `NULL` the values of the `yticks` are used. It is recommended that 1 is in the `yvalues` if the `plottype = "hazard"` and 0 if the `plottype = "loghazard"`. In addition to this the `yvalues` should be a subset of `yticks`.

plottype

A string that defines the scale for the y-axis. Either `"hazard"` for the relative hazard with log-scale or `"loghazard"` for the logarithm of the relative hazard with linear scale.

na.rm

By default `TRUE` and only complete cases are plotted. Complete cases are cases that have information on all covariates that are used in fitting the model. If `FALSE` all available cases for each variable are plotted.

draw

By default `TRUE` and a rank-hazard plot and a summary are printed. If `FALSE`, no output is provided unless `return = TRUE`.

return

By default `FALSE` and `Value` is not returned. Used with the argument `coxphobj` or `cphobj`. If `TRUE`, `x`, `xp`, `refvalues` and `confinterval` are returned as a list. See section `Value`.

col

A vector that defines the colours of the lines and the points. If the vector is shorter than the number of the covariates to be plotted, the values are repeated. See documentation for `par`.

lwd

A vector that defines the widths of the lines. If the vector is shorter than the number of the covariates to be plotted, the values are repeated. See documentation for `par`.

lty

A vector that defines the types of the lines. If the vector is shorter than the number of the covariates to be plotted, the values are repeated. See documentation for `par`.

pch

A vector that defines the characters of the points. If the vector is shorter than the number of the covariates to be plotted, the values are repeated. See documentation for `points`.

bg

A vector that defines the fill colour of the point. Available only for point characters `21:25`. If the vector is shorter than the number of the covariates to be plotted, the values are repeated. By default `"transparent"`. See documentation for `points`.

pt.lwd

A vector that defines the line width for the drawing symbols. If the vector is shorter than the number of the covariates to be plotted, the values are repeated. See documentation for `lwd` in `points`.

cex

A vector that defines the size of the points. If the vector is shorter than the number of covariates to be plotted, the values are repeated. See documentation for `plot.default`.

...

Other arguments to be passed to `plot` and `lines` commands. For example `main`.

B.4 Details

The function `rankhazardplot` receives a `coxph` (package `survival`) object or a `cph` (package `rms`) object as an argument and creates a rank-hazard plot of the covariates. The reference points for the relative hazards and legend texts can be provided as optional arguments. Plotting parameters such as, lwd, lty, col and pch are passed to the plotting commands.

Rank-hazard plots visualize the relative importance of covariates in a proportional hazards model. The key idea is to rank the covariate values and plot the relative hazard as a function of ranks scaled to interval [0,1]
. The relative hazard is the hazard plotted in respect to the reference hazard, which can be e.g. the hazard related to the median of the covariate.

The labels on the x-axis show the minimum, the quartiles and the maximum of each covariate. These are real observations. If the quantile would be determined by a mean of two observations, the smaller value is choosed to be the quantile. However, if the number of the obervations is even, the default reference point is not necessarily a real observation as it is calculated as a mean of two middle observations. Hence, the median shown on the x-axis and the reference point can differ even when the default is used.

Predictions are computed by `predict.coxph`, when the function is called with the argument `coxphobj` and by `predict.rms`, when the function

is called with the argument `cphobj`. Consequently, relative hazards are available for models that are supported by `predict.coxph` or `predict.rms`. For example the `pspline` transforms are supported by `predict.coxph` but not by `predict.rms`.

The upper and lower confidence limits of the Cox regression coefficients are used to calculate the confidence intervals for the relative hazards. Confidence intervals are only supported for original covariates (same as in the data), simply transformed covariates (e.g. `log`) and factors.

Rank-hazard plots can be used to visualize time-dependent models. In that case plotting can be made using `coefs` that are Cox regression coefficients of the time-dependent model. The data matrix `x` contains the values of covariates at some specific time. It is also possible to make a rank-hazard plot by the argument `xp`. In that case both `x` and `xp` must be selected so that they contain information only at some specific time. Third way to make the rank-hazard plot is to use the time-dependent model and give the values of the covariates at some specific time as the argument `data`. If the purpose is to compare relative hazards at different times, it is recommended that the same reference points are used in every plot. For example the medians of every covariate at the first measurement.

B.5 Value

If `return = TRUE` the function returns a list that contains all the information needed to draw a rank-hazard plot and confidence intervals. The list contains:

x
: A data frame that contains the covariate data.

xp
: A data frame that contains the centered predictions for all covariates in the model. Calculated by `predict.coxph` or `predict.rms`.

refvalues
: A vector that contains the centered predictions that are calculated using the reference point defaults for the covariates that don't have a given reference point. Calculated by `predict.coxph` or `predict.rms`.

confinterval
: A list that contains covariate data as a data frame `x`, predictions by terms as a data frame `xp`, reference values as a vector `refvalues`, a lower confidence interval for the predictions as a data frame `low`, a lower confidence interval for the reference values as a vector `lowrefvalues`, an upper confidence interval for the predictions as a data frame `upp` and an upper confidence interval for the reference values as a vector `upprefvalues`.

The covariates for which the confidence intervals are provided are original (same as in the data), simply transformed (e.g. `log`) and factors. The predictions are calculated as a product of `coefs` and `x`. The upper and lower confidence limits of the Cox regression coefficients are used to calculate the confidence intervals for the relative hazards. NB: values aren't centered and for that reason e.g. `xp` and `confinterval$xp` are not the same.

B.6 Author(s)

Juha Karvanen <juha.karvanen@iki.fi>, Nanni Koski

B.7 References

Karvanen J., Harrell F. E., Jr. 2009 Visualizing covariates in proportional hazards model. *Statistics in Medicine*, **28**, 1957–1966.

B.8 See also

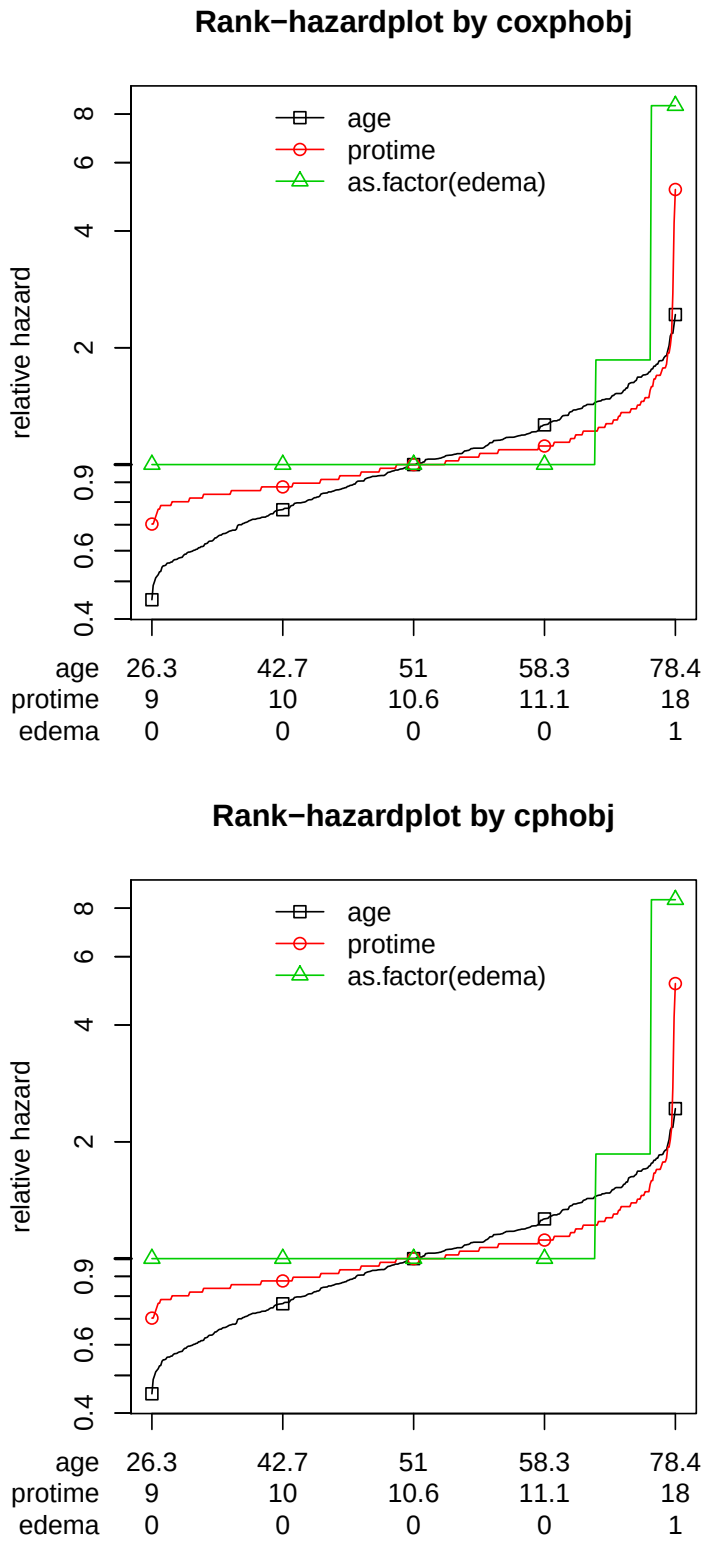
`coxph`, `cph`, `predict.coxph`, `predict.rms`

C Dokumentaation esimerkit ja tulosteet

```
> library(survival)
> library(rms)
>
> data(pbc)
> # new status variable
> pbc$statusbin <- ifelse(pbc$status==0, 0, NA)
> pbc$statusbin <- ifelse(pbc$status==2, 1, pbc$statusbin)
```

C.1 Kuva 11: Piirtäminen malliobjektia käyttäen

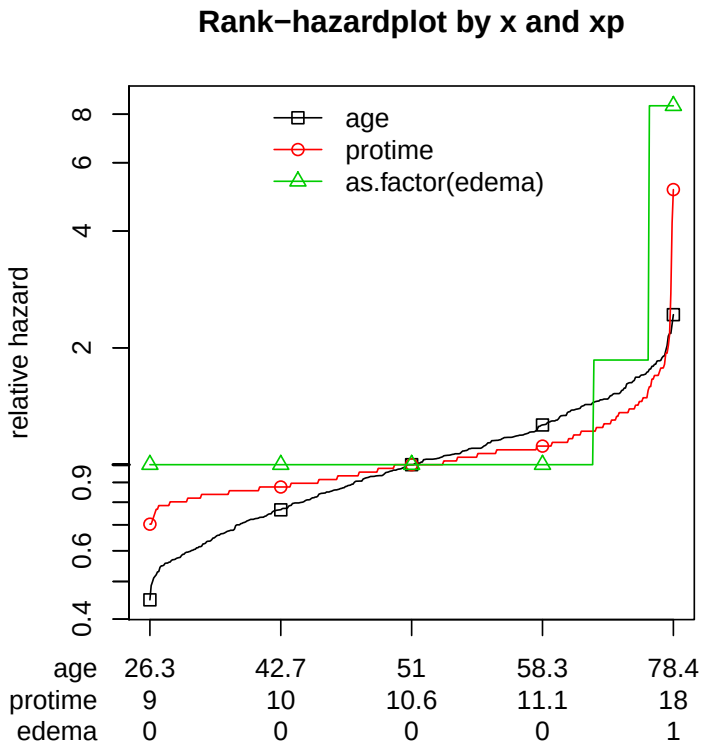
```
> ### different ways to make a rank-hazard plot ###
> par(mar = c(4, 5, 4, 2) + 0.1);  par(mfrow = c(2, 1))
>
> coxmodel1 <- coxph(Surv(time, statusbin) ~ age + protime +
+     as.factor(edema), data = pbc, x = TRUE)
> rankhazardplot(coxmodel1, data = pbc,
+     main = "Rank-hazardplot by coxphobj")


Y-axis range:  0.448 8.41

Relative hazards for each covarite:
                  Min. 1st Qu. Median 3rd Qu. Max.
age              0.448   0.766      1    1.27 2.44
protime          0.703   0.876      1    1.12 5.12
as.factor(edema) 1.000   1.000      1    1.00 8.41


> dd <- datadist(pbc)
> options(datadist = 'dd')

> cmodel1 <- cph(Surv(time, statusbin) ~ age + protime +
+     as.factor(edema), data = pbc, x = TRUE)
> rankhazardplot(cmodel1, data = pbc,
+     main = "Rank-hazardplot by cphobj")


Y-axis range:  0.448 8.41

Relative hazards for each covarite:
                  Min. 1st Qu. Median 3rd Qu. Max.
age              0.448   0.766      1    1.27 2.44
protime          0.703   0.876      1    1.12 5.12
as.factor(edema) 1.000   1.000      1    1.00 8.41
```



Kuva 11: Esimerkit riskitiheyskuvion piirtämisestä `coxphobj`- ja `cphobj`-argumentteja käyttäen.

C.2 Kuva 12: Piirtäminen xp-argumenttia käyttäen

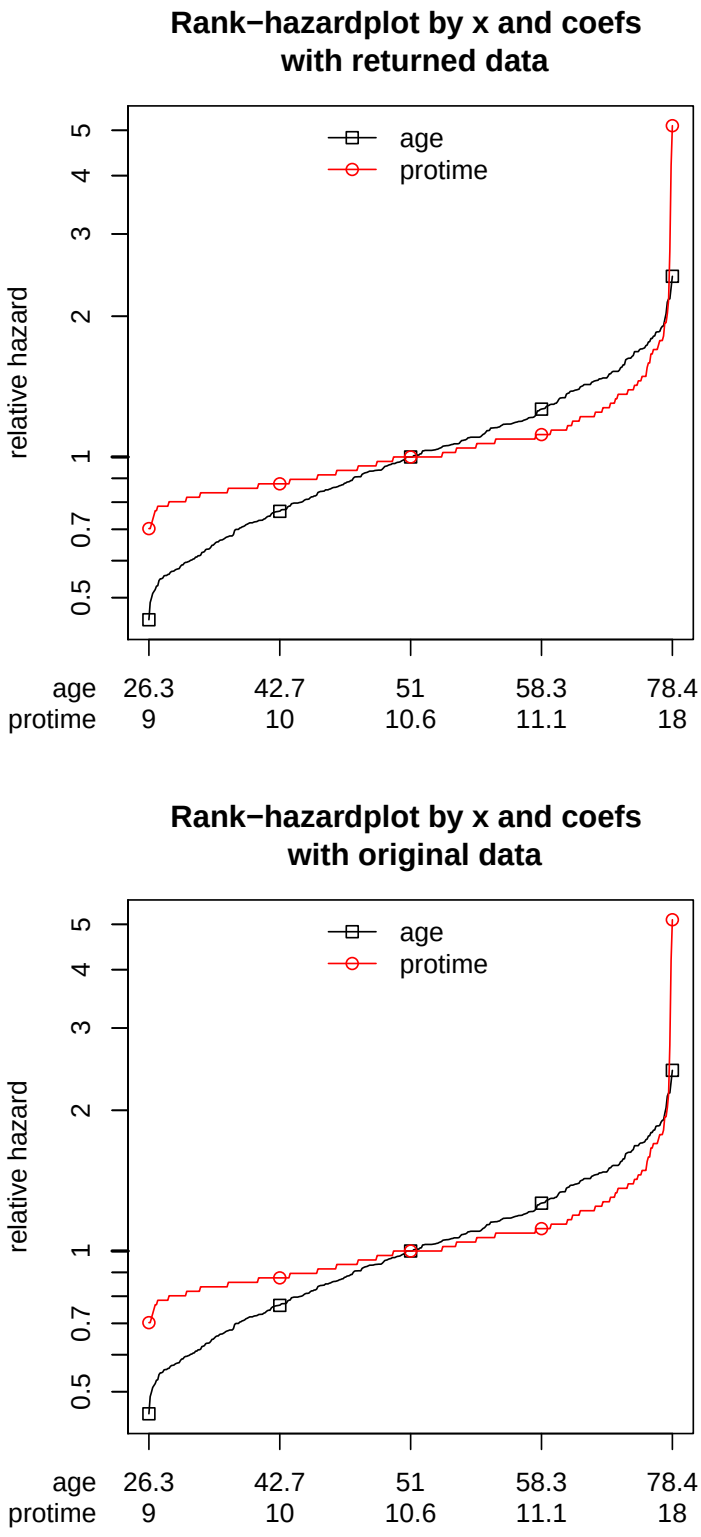
```
> output1 <- rankhazardplot(coxmodel1, data = pbc, draw = FALSE,
+     return = TRUE)
> rankhazardplot(x = output1$x, xp = output1$xp,
+     refvalues = output1$refvalues,
+     main = "Rank-hazardplot by x and xp")


Y-axis range:  0.448 8.41

Relative hazards for each covarite:
                   Min. 1st Qu. Median 3rd Qu. Max.
age               0.448   0.766      1    1.27 2.44
protime           0.703   0.876      1    1.12 5.12
as.factor(edema) 1.000   1.000      1    1.00 8.41
```



Kuva 12: Esimerkki riskitiheyskuvion piirtämisestä xp-argumenttia käyttäen.

C.3 Kuva 13: Piirtäminen `coefs`-argumenttia käyttäen

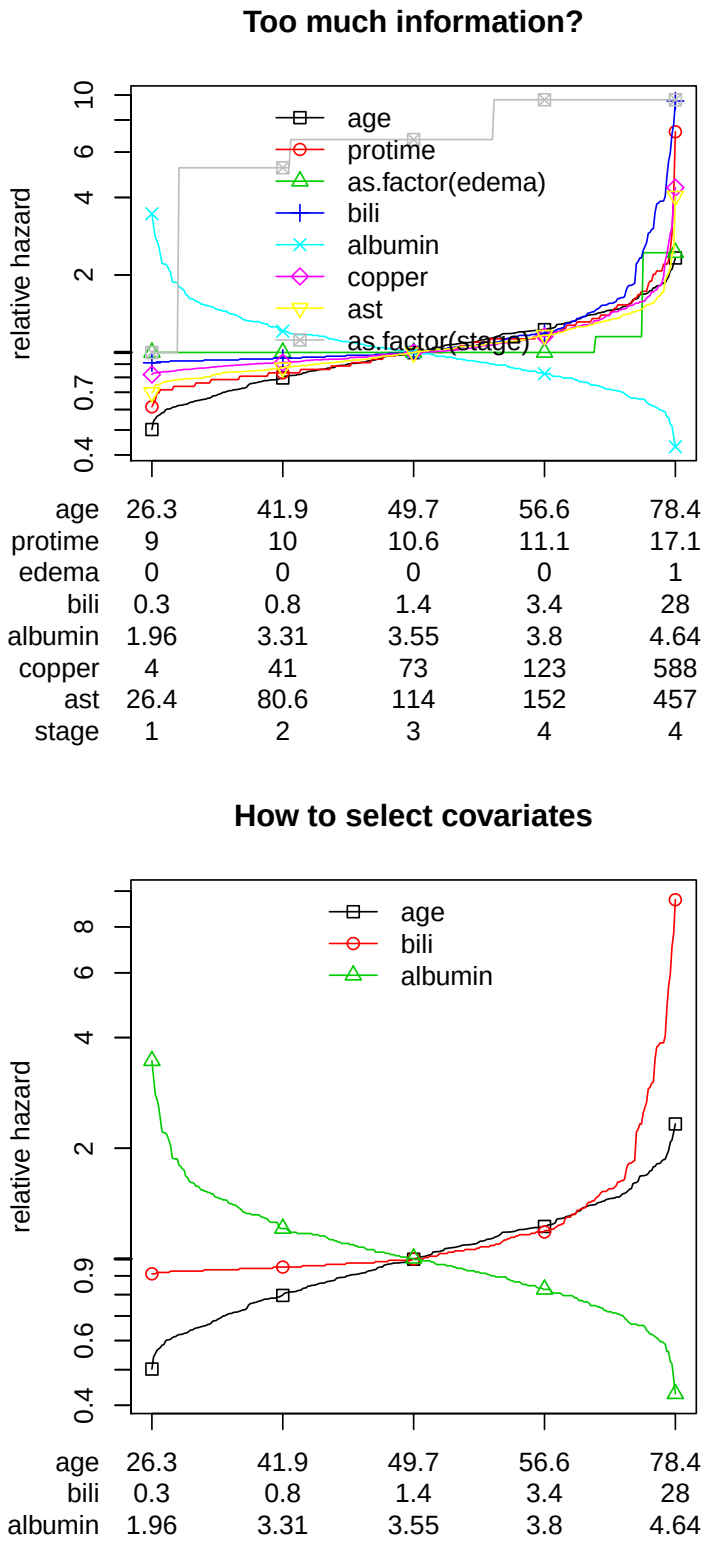
```
> par(mar = c(4, 5, 4, 2) + 0.1); par(mfrow = c(2, 1))
>
> rankhazardplot(x = output1$x[1:2], coefs = coxmodel1$coef[1:2],
+    main = "Rank-hazardplot by x and coefs
+    with returned data")


Y-axis range:  0.448 5.12

Relative hazards for each covarite:
         Min. 1st Qu. Median 3rd Qu. Max.
age     0.448   0.766      1    1.27 2.44
protime 0.703   0.876      1    1.12 5.12


> rankhazardplot(x = pbc[c("age","protime")],
+     coefs = coxmodel1$coef[1:2],
+     main = "Rank-hazardplot by x and coefs
+     with original data")


Y-axis range:  0.448 5.12

Relative hazards for each covarite:
         Min. 1st Qu. Median 3rd Qu. Max.
age     0.448   0.766      1    1.27 2.44
protime 0.703   0.876      1    1.12 5.12
```



Kuva 13: Esimerkit riskitiheyskuvion piirtämisestä `coefs`-argumenttia käyttäen. Data sisältyy argumenttiin `x`, ja data voi olla alkuperäinen tai `rankhazardplot`-funktion palauttama.

C.4 Kuva 14: Piirrettävien kovariaattien valinta

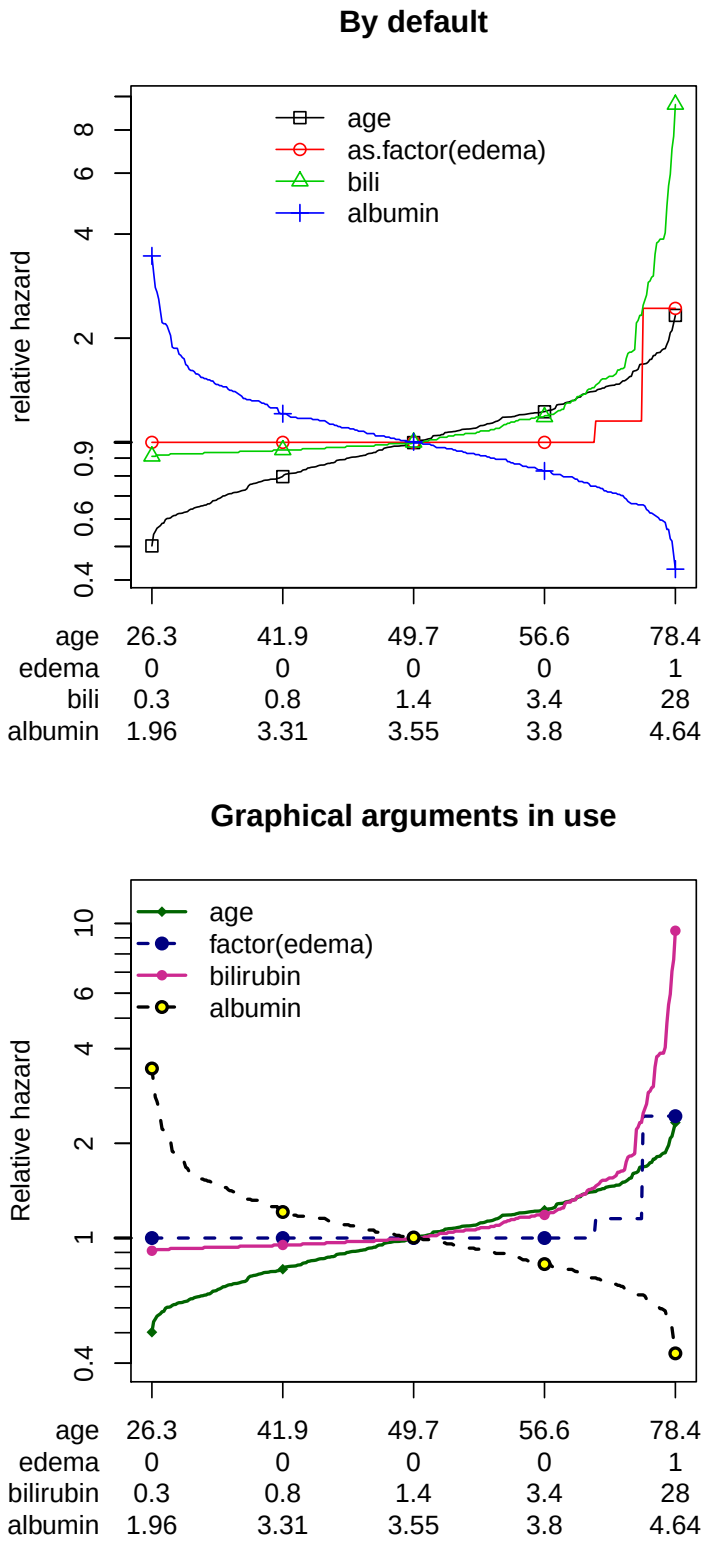
```
> ### selecting covariates ###
> par(mfrow = c(2, 1))
> coxmodel2 <- coxph(Surv(time, statusbin) ~ age + protime +
+     as.factor(edema) + bili + albumin + copper + ast +
+     as.factor(stage), data = pbc, x = TRUE)
> par(mar = c(9, 5, 4, 2) + 0.1)
> rankhazardplot(coxmodel2, data = pbc,
+     main = "Too much information?")


Y-axis range:  0.43 9.58

Relative hazards for each covarite:
                  Min. 1st Qu. Median 3rd Qu. Max.
age              0.502   0.798   1.00    1.23 2.33
protime          0.615   0.833   1.00    1.16 7.20
as.factor(edema) 1.000   1.000   1.00    1.00 2.44
bili             0.911   0.951   1.00    1.19 9.48
albumin          0.430   0.826   1.00    1.21 3.46
copper           0.821   0.913   1.00    1.15 4.37
ast              0.699   0.872   1.00    1.17 4.05
as.factor(stage) 1.000   5.220   6.71    9.58 9.58


> par(mar = c(4, 5, 4, 2) + 0.1)
> rankhazardplot(coxmodel2, data = pbc, select = c(1, 4, 5),
+     main = "How to select covariates")


Y-axis range:  0.43 9.48

Relative hazards for each covarite:
         Min. 1st Qu. Median 3rd Qu. Max.
age     0.502   0.798      1    1.23 2.33
bili    0.911   0.951      1    1.19 9.48
albumin 0.430   0.826      1    1.21 3.46
```



Kuva 14: Esimerkki piirrettävien kovariaattinen valinnasta. Liian monta suhteellisen riskitiheyden jakaumaa esitettynä samassa kuvassa voi heikentää kuvan selkeyttä. Valinta on helpointa toteuttaa `select`-argumenttia käyttäen.

C.5 Kuva 15: Graafisten ominaisuuksien muokkaus

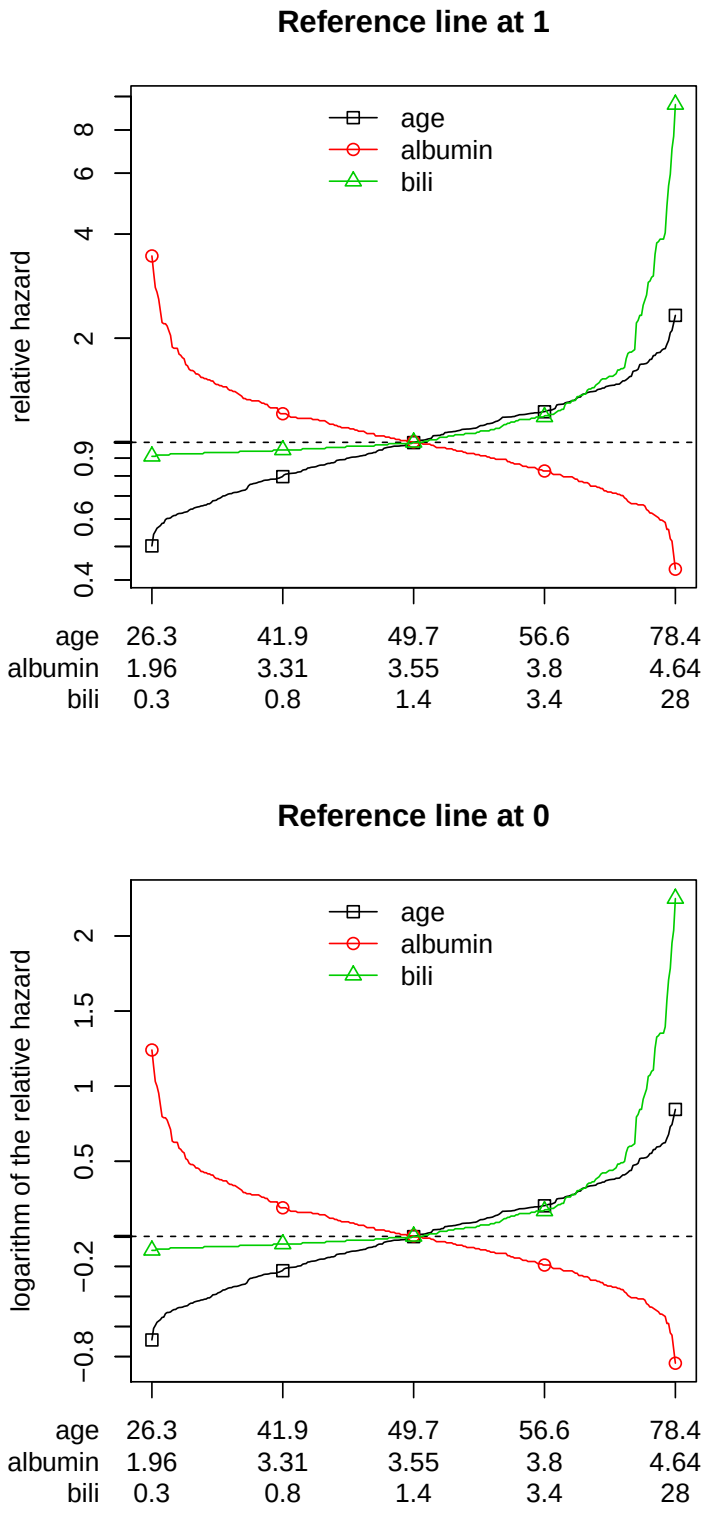
```
> ### using graphical arguments ###
>
> # Compare the two following plots
> par(mar = c(5, 5, 4, 2) + 0.1); par(mfrow = c(2, 1))
> rankhazardplot(coxmodel2, select = c(1, 3, 4, 5), data = pbc,
+     main = "By default")


Y-axis range:  0.43 9.48

Relative hazards for each covarite:
                  Min. 1st Qu. Median 3rd Qu. Max.
age              0.502   0.798      1    1.23 2.33
as.factor(edema) 1.000   1.000      1    1.00 2.44
bili             0.911   0.951      1    1.19 9.48
albumin          0.430   0.826      1    1.21 3.46


> # X11() # If the plot is too small, legend is not printed
>     #in the correct place
> par(mar = c(5, 5, 4, 2) + 0.1)
> rankhazardplot(coxmodel2, select = c(1, 3, 4, 5), data = pbc,
+     legendlocation = "topleft", ylim = c(0.4, 12),
+     ylab = "Relative hazard", yvalues = c(0.4, 1, 2, 4, 6, 10),
+     yticks = c(seq(0.4, 1, by = 0.1), 2:10),
+     main = "Graphical arguments in use",
+     axistext = c("age", "edema", "bilirubin", "albumin"),
+     legendtext = c("age", "factor(edema)", "bilirubin", "albumin"),
+     col = c("darkgreen","navyblue","maroon3",1), pch = 18:21,
+     lwd = 2, lty = c(1, 2), cex = 0.9, bg = "yellow", pt.lwd = 2)


Y-axis range:  0.4 12

Relative hazards for each covarite:
                Min. 1st Qu. Median 3rd Qu. Max.
age            0.502   0.798      1    1.23 2.33
factor(edema) 1.000   1.000      1    1.00 2.44
bilirubin      0.911   0.951      1    1.19 9.48
albumin        0.430   0.826      1    1.21 3.46
```



Kuva 15: Esimerkki graafisten argumenttien käyttämisestä. Vertailussa oletusarvoilla piirrettävä kuva (ylhäällä) ja kuva, johon on säädetty useita graafisia ominaisuuksia.

C.6 Kuva 16: Referenssin korostaminen

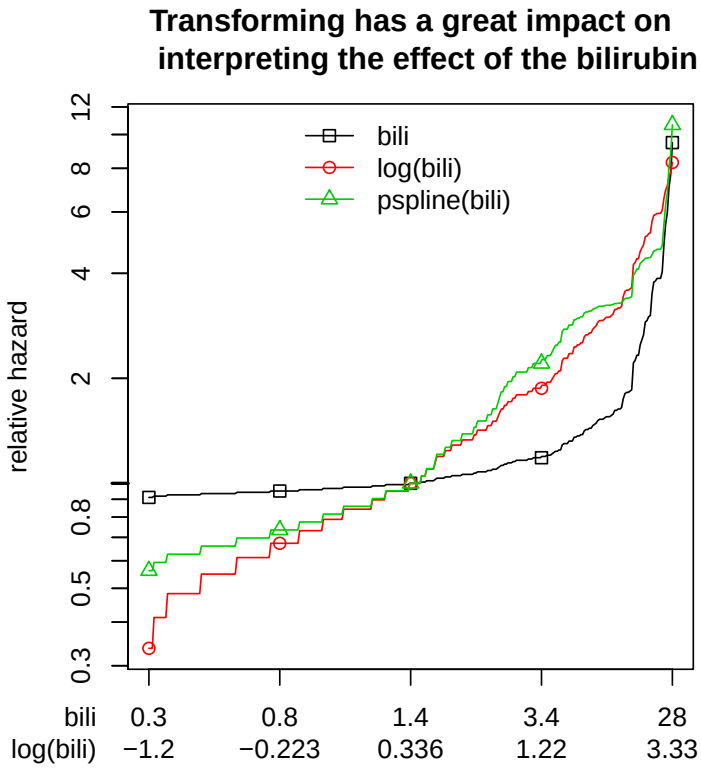
```
> # reference line for hazard and loghazard #
> par(mar = c(5, 5, 4, 2) + 0.1); par(mfrow = c(2, 1))
> rankhazardplot(coxmodel2, select = c(1, 5, 4), data = pbc,
+     refline = TRUE, plottype = "hazard",
+     main = "Reference line at 1")


Y-axis range:  0.43 9.48

Relative hazards for each covarite:
         Min. 1st Qu. Median 3rd Qu. Max.
age     0.502   0.798      1    1.23 2.33
albumin 0.430   0.826      1    1.21 3.46
bili    0.911   0.951      1    1.19 9.48


> rankhazardplot(coxmodel2, select = c(1, 5, 4), data = pbc,
+     refline = TRUE, plottype = "loghazard",
+     main = "Reference line at 0")


Y-axis range:  -0.844 2.25

Logarithm of the relative hazards for each covarite:
          Min. 1st Qu.    Median 3rd Qu.  Max.
age     -0.689 -0.2250 -1.04e-16   0.205 0.845
albumin -0.844 -0.1910 -1.73e-16   0.191 1.240
bili    -0.093 -0.0507  0.00e+00   0.175 2.250
```



Kuva 16: Esimerkki referenssin osoittamisesta (katkoviiva) argumenttia `refline` käyttäen. Referenssi suhteelliselle riskitiheydelle on 1 ja sen logaritmille 0.

C.7 Kuva 17: Muunnosten piirtäminen samaan kuvaan

```
> ### comparing covariates from different models ###
>
> # transforms #
> # same model as coxmodel2, only bilirubin is transformed by log
> logmodel <- coxph(Surv(time, statusbin) ~ age + protime +
+     as.factor(edema) + log(bili) + albumin + copper + ast +
+     as.factor(stage), data = pbc, x = TRUE)
> # same model as coxmodel2, only a pspline is fitted to bilirubin
> coxspline <- coxph(Surv(time, statusbin) ~ age + protime +
+     as.factor(edema) + pspline(bili) + albumin + copper +
+     ast + as.factor(stage), data = pbc, x = TRUE)
>
> outputcox <- rankhazardplot(coxmodel2, data = pbc, return = TRUE,
+     draw = FALSE)
> outputlog <- rankhazardplot(logmodel, data = pbc, return = TRUE,
+     draw = FALSE)
> outputspline <- rankhazardplot(coxspline, data = pbc, return = TRUE,
+     draw = FALSE)
>
> xlog <- data.frame(outputcox$x["bili"], log(outputlog$x["bili"]),
+     outputspline$x["bili"])
> xp <- data.frame(outputcox$xp["bili"], outputlog$xp["log(bili)"],
+     outputspline$xp["pspline(bili)"])
> ref <- c(outputcox$ref["bili"], outputlog$ref["log(bili)"],
+     outputspline$ref["pspline(bili)"])
> par(mar = c(3, 5, 4, 2) + 0.1); par(mfrow = c(1, 1))
> rankhazardplot(x = xlog, xp = xp, refvalues = ref,
+     legendtext = c( "bili","log(bili)", "pspline(bili)"),
+     axistext = c("bili", "log(bili)", "bili"),
+     main = "Transforming has a great impact on
+     interpreting the effect of the bilirubin")


Y-axis range:  0.336 10.7

Relative hazards for each covarite:
               Min. 1st Qu. Median 3rd Qu.  Max.
bili          0.911   0.951      1    1.19  9.48
log(bili)     0.336   0.673      1    1.90  8.32
pspline(bili) 0.562   0.735      1    2.25 10.70
```

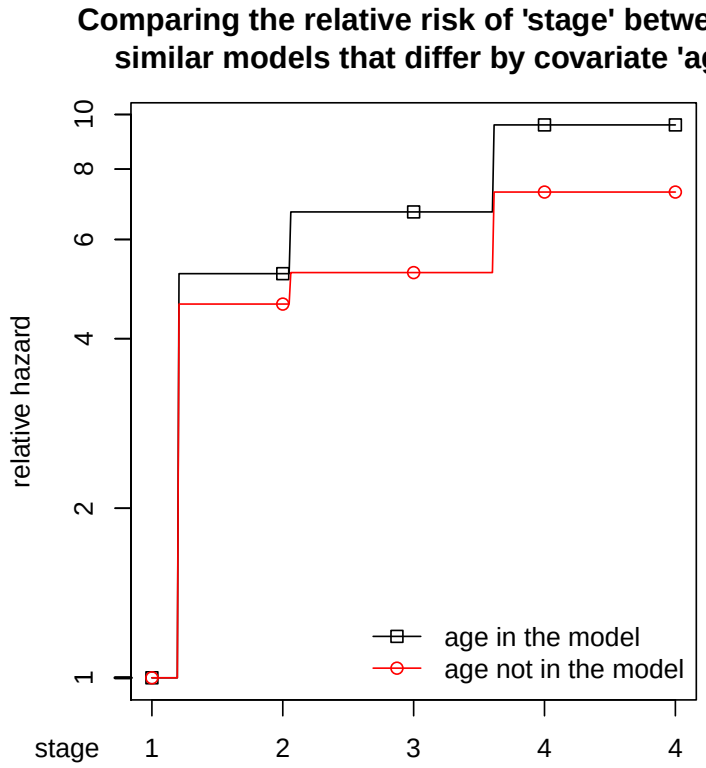


Kuva 17: Esimerkki muunnoksen käyttämisen vaikutuksesta suhteellisen riskitiheyden jakaumaan. Kuvassa on bilirubiini-muuttujan alkuperäinen, logaritmi-muunnettu ja spline-sovitettu suhteellisen riskitiheyden jakauma kolmesta eri mallista, joissa on muuten samat kovariaatit.

C.8 Kuva 18: Faktoreiden piirtäminen samaan kuvaan

```
> # factors #
> # same model as coxmodel2, only age is left out
> # how does it affect relative hazards for stage?
> coxmodel3 <- coxph(Surv(time, statusbin) ~ protime +
+     as.factor(edema) + log(bili) + albumin + copper +
+     ast + as.factor(stage), data = pbc, x = TRUE)
> outputcox2 <- rankhazardplot(coxmodel3, data = pbc,
+     return = TRUE, draw = FALSE)
> xp <- data.frame(outputcox$xp["as.factor(stage)"],
+     outputcox2$xp["as.factor(stage)"])
> x <- data.frame(outputcox$x["stage"],outputcox2$x["stage"])
> ref <- c(outputcox$ref["as.factor(stage)"],
+     outputcox2$ref["as.factor(stage)"])
>
> par(mar = c(2, 5, 4, 2) + 0.1)
> rankhazardplot(x = x, xp = xp, refvalues = ref,
```

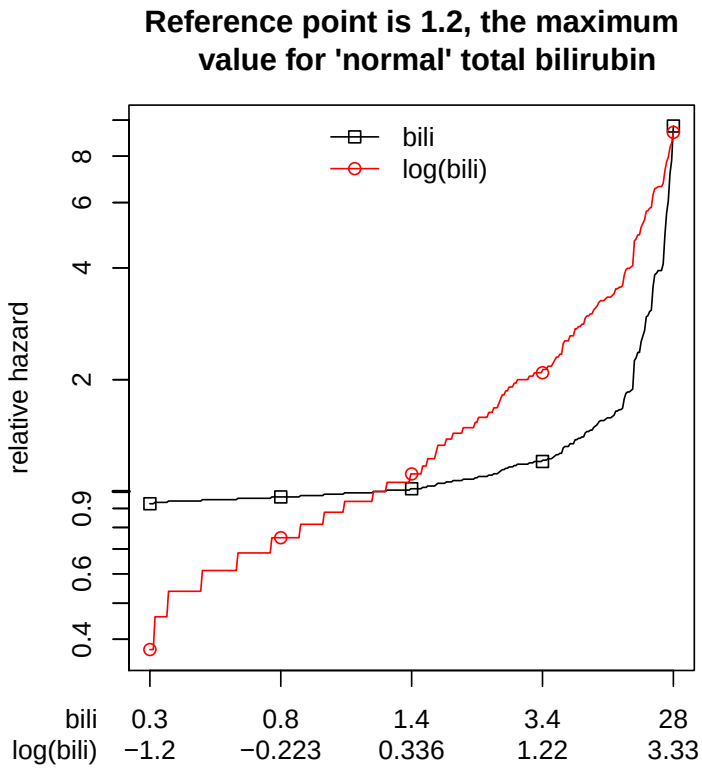
```
+     legendlocation = "bottomright",
+     legendtext = c("age in the model","age not in the model"),
+     axistext = c("stage","stage"),
+     main = "Comparing the relative risk of 'stage' between
+     similar models that differ by covariate 'age'")

Y-axis range:  1 9.58

Relative hazards for each covarite:
                     Min. 1st Qu. Median 3rd Qu. Max.
age in the model        1    5.22   6.71    9.58 9.58
age not in the model    1    4.61   5.24    7.28 7.28
```



Kuva 18: Esimerkki mallista poistetun ikä-muuttujan vaikutuksesta sairauden aste -kovariaatin suhteellisen riskitiheyden jakaumaan sekä eri malleissa olevien faktoreiden piirtämisestä samaan kuvaan.

C.9 Kuva 19: Vertailukohdan vaihtaminen

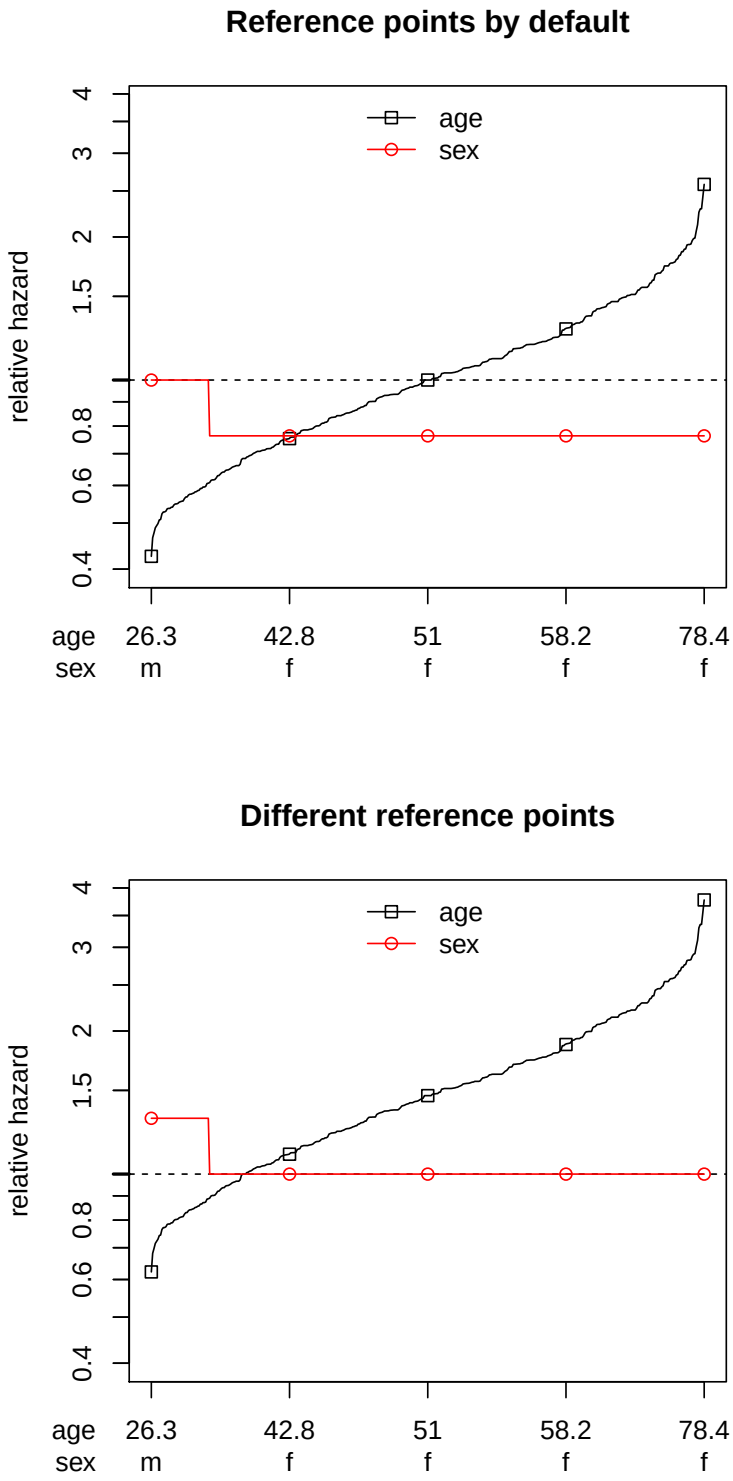
```
> ### changing reference points ###
>
> # with no model object #
> coefs <- c(coxmodel2$coef["bili"],logmodel$coef["log(bili)"])
> par(mar = c(3, 5, 4, 2) + 0.1)
> rankhazardplot(x = xlog[1:2], coefs = coefs[1:2],
+     refpoints = c(1.2, log(1.2)),
+     legendtext = c("bili", "log(bili)"),
+     main = "Reference point is 1.2, the maximum
+     value for 'normal' total bilirubin")

Y-axis range:  0.375 9.64

Relative hazards for each covarite:
           Min. 1st Qu. Median 3rd Qu. Max.
bili      0.927   0.967   1.02    1.21 9.64
log(bili) 0.375   0.751   1.12    2.12 9.27
```



Kuva 19: Esimerkki vertailukohdan vaihtamisesta sisällöllisesti merkitykselliseksi. Vertailukohtaa pienemmillä arvoilla bilirubiiniarvoa pidetään normaalina, vertailukohtaa suuremmilla arvoilla kohonneena.

C.10 Kuva 20: Vertailukohdan vaihtaminen faktorille

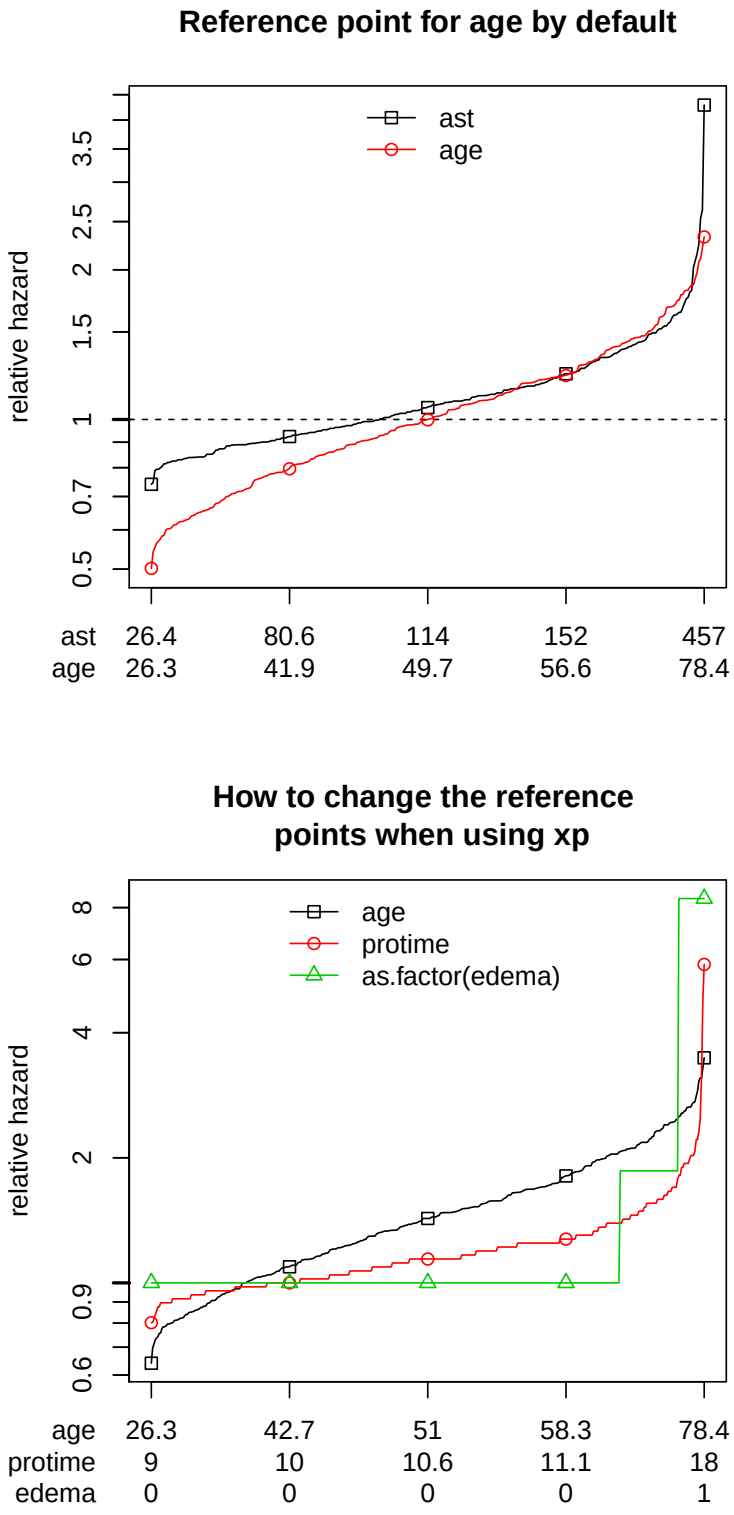
```
> # factors with non-numerical levels#
> coxmodel4 <- coxph(Surv(time, statusbin) ~ age + sex, data = pbc,
+     x = TRUE)
> par(mfrow = c(2, 1))
> rankhazardplot(coxmodel4, data = pbc,
+     main = "Reference points by default",
+     refline = TRUE, ylim = c(0.4, 3.8))


Y-axis range:  0.4 3.8

Relative hazards for each covarite:
     Min. 1st Qu. Median 3rd Qu. Max.
age 0.426   0.754  1.000   1.280 2.58
sex 0.763   0.763  0.763   0.763 1.00


> rankhazardplot(coxmodel4, refpoints = c(40, "f"), data = pbc,
+     main = "Different reference points",
+     refline = TRUE,ylim = c(0.4, 3.8))


Y-axis range:  0.4 3.8

Relative hazards for each covarite:
     Min. 1st Qu. Median 3rd Qu. Max.
age 0.622     1.1   1.46    1.88 3.77
sex 1.000     1.0   1.00    1.00 1.31
```



Kuva 20: Havainnollistus vertailukohdan vaihtamisesta jatkuvalla muuttujalla sekä ei-numeerisella faktorilla. Suhteellisen riskitiheyden jakauman kuvaaja leikkaa referenssisuoran vertailukohdassa.

C.11 Kuva 21: Vertailukohdan vaihtaminen osalle kovariaateista sekä xp-argumenttia käyttäen

```
> par(mfrow = c(2, 1))
> # with select argument
> # changing only part of reference points #
> rankhazardplot(coxmodel2, data = pbc, select = c(7, 1),
+     refpoints = c(100, NA), ylim = c(0.5, 4.3), refline = TRUE,
+     main = "Reference point for age by default")


Y-axis range:  0.5 4.3

Relative hazards for each covarite:
     Min. 1st Qu. Median 3rd Qu. Max.
ast 0.741   0.924   1.06    1.24 4.29
age 0.502   0.798   1.00    1.23 2.33


>
> # using xp argument #
> output2 <- rankhazardplot(coxmodel1, refpoints = c(40, 10,0),
+     data = pbc, draw = FALSE, return = TRUE)
> rankhazardplot(x = output2$x, xp = output2$xp,
+     refvalues = output2$refvalues,
+     main = "How to change the reference
+     points when using xp")


Y-axis range:  0.641 8.41

Relative hazards for each covarite:
                  Min. 1st Qu. Median 3rd Qu. Max.
age              0.641    1.09   1.43    1.81 3.48
protime          0.802    1.00   1.14    1.27 5.84
as.factor(edema) 1.000    1.00   1.00    1.00 8.41
```



Kuva 21: Ylhäällä esimerkki vertailukohdan vaihtamisesta vain toiselle muuttujalle. Alhaalla esimerkki vertailukohtien vaihtamisesta, kun kuva piirretään argumenttia `xp`-käyttäen.

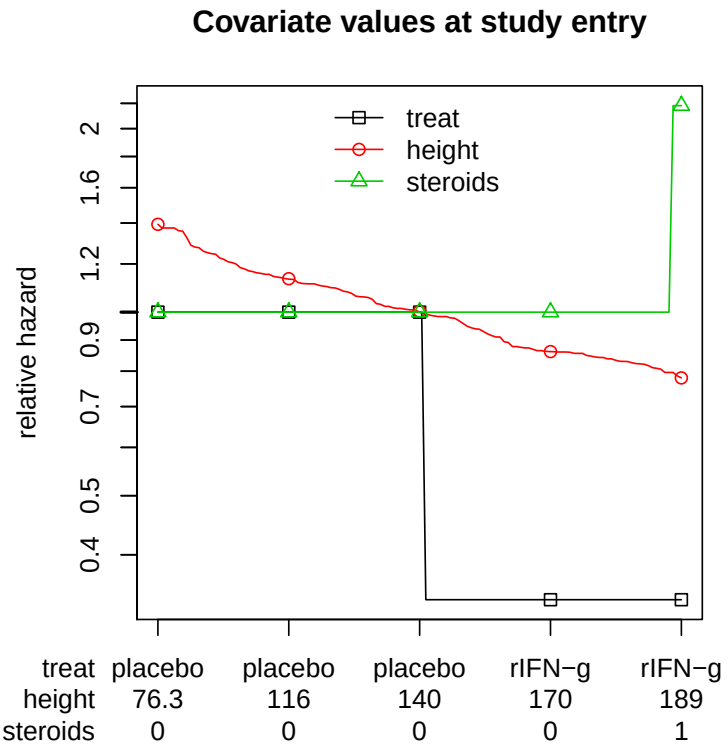
C.12 Kuva 22: Suhteellisen riskitiheyden jakauma toistuvien sairastumisten mallille

```
> ### data in start-stop format ###
> data(cgd)
> timemodel <- coxph(Surv(tstart, tstop, status) ~ treat + height +
+     steroids + cluster(id), data = cgd, x = TRUE)
> # steroids and height are in the model only to make
> # the example plot more interesting
> par(mar = c(4, 5, 4, 2) + 0.1); par(mfrow = c(1, 1))
> rankhazardplot(timemodel, data = cgd[cgd$enum == 1,],
+    axistextposition  = -0.12, select = 1:3,
+     main = "Covariate values at study entry")


Y-axis range:  0.338 2.18

Relative hazards for each covarite:
          Min. 1st Qu. Median 3rd Qu. Max.
treat    0.338   0.338      1    1.00 1.00
height   0.780   0.861      1    1.13 1.39
steroids 1.000   1.000      1    1.00 2.18
```

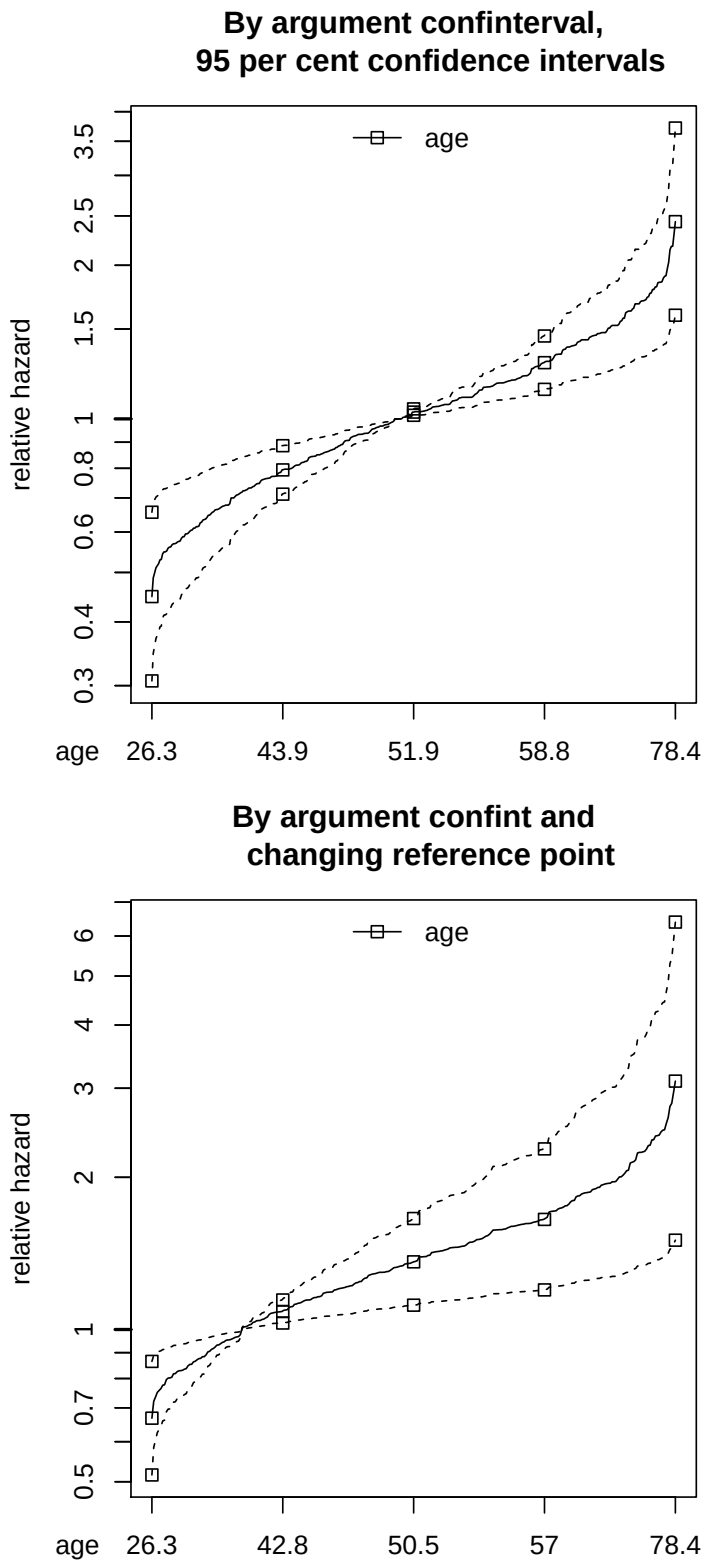
C.13 Kuva 23: Luottamusvälit

```
> ### confidence intervals ###
> par(mar = c(2, 5, 4, 2) + 0.1); par(mfrow = c(2, 1))
> rankhazardplot(confinterval = output1$conf,
+     main = "By argument confinterval,
+     95 per cent confidence intervals")


Y-axis range:  0.306 3.72

Relative hazards for each covarite:
     Min. 1st Qu. Median 3rd Qu. Max.
age 0.448   0.795   1.03    1.29 2.44

Relative hazards for the confidence intervals of each covariate:
         Min. 1st Qu. Median 3rd Qu. Max.
Low_age 0.656   0.886   1.02    1.14 1.60
Upp_age 0.306   0.713   1.05    1.46 3.72
```

```
> rankhazardplot(coxmodel2, data = pbc, confint = TRUE, select = 1,
+     refpoint = 40, main = "By argument confint and
+     changing reference point")


Y-axis range:  0.516 6.39

Relative hazards for each covarite:
     Min. 1st Qu. Median 3rd Qu. Max.
age 0.668    1.09   1.36    1.65  3.1

Relative hazards for the confidence intervals of each covariate:
         Min. 1st Qu. Median 3rd Qu. Max.
Low_age 0.865    1.03   1.12    1.20 1.50
Upp_age 0.516    1.15   1.66    2.28 6.39
```



Kuva 22: Esimerkki riskitiheyskuvion käyttämisestä start-stop–muodossa olevan aineiston esittämiseen. Aineistossa on mitattu aikaa tulehduksien esiintymisen välillä, jolloin kullekin yksilölle tulehduksia voi tulla useita. Kuvan suhteellisen riskitiheyden jakaumat on määritetty käyttäen kultakin henkilöltä tutkimuksen alussa mitattuja arvoja.



Kuva 23: Esimerkki luottamusvälien piirtämisestä (ylhäällä) ja vertailukohdan vaihtamisen vaikutuksesta luottamusväleihin (alhaalla). Luottamusvälit risteävät vertailukohdassa. Kuvaajat on piirretty eri malleista, jolloin ikäjakaumat eivät ole täysin samat.

D Koodit rankhazard-paketissa

D.1 rankhazardplot

```
rankhazardplot <- function(...) UseMethod("rankhazardplot")
```

D.2 rankhazardplot.default

```
rankhazardplot.default <- function ( \label{def1}
    x, coefs = NULL, xp = NULL, refvalues = NULL, refpoints = NULL,
    confinterval = NULL, select = 1, legendtext = NULL,
    axistext = NULL, legendlocation = "top", axistextposition = -0.1,
    reftick = TRUE, refline = FALSE, refline.col = 1, refline.lwd = 1,
    refline.lty = 2, ylab = NULL, ylim = NULL, yticks = NULL,
    yvalues = NULL, plottype = "hazard", na.rm = TRUE,
    col = NULL, lwd = 1, lty = 1, pch = NULL,
    cex = 1, bg = "transparent", pt.lwd = 1, ...)
{
    if (!is.null(confinterval)){
        x <- confinterval$x
        if (na.rm) x <- na.omit(x)
        x <- confinterval$x[select]
        xp <- confinterval$xp
        if (na.rm) xp <- na.omit(xp)
        xp <- confinterval$xp[select]
        refvalues <- confinterval$refvalues[select]
    }

```

```
    if (na.rm) x <- na.omit(x)

    if (na.rm & !is.null(xp)) xp <- na.omit(xp)

    n <- dim(x)[1] # number of observations \label{def2}
    m <- dim(x)[2] # number of covariates

    if (!identical(plottype, "hazard") & !identical(plottype, "loghazard"))
        stop("Unknown plottype")
    if (is.null(xp) & is.null(coefs))
        stop("Either coefs or xp must be provided.")
    if (is.null(refvalues) & !is.null(xp))
        stop("When xp is given, also refvalues are required.")

    if (is.null(refvalues) & is.null(refpoints)){
        refpoints <- apply(x, 2, median, na.rm = TRUE)
        refvalues <- coefs*refpoints
    }

    if (is.null(refvalues) & !is.null(refpoints))
        refvalues <- coefs*refpoints

    if (is.null(xp))
        xp <- as.data.frame(t(coefs * t(x)))

    lwd <- rep(lwd, length.out = m)
    lty <- rep(lty, length.out = m)
```

```
    cex <- rep(cex, length.out = m)
    bg <- rep(bg, length.out = m)
    pt.lwd <- rep(pt.lwd, length.out = m)

    if (is.null(pch)){ pch <- seq(0, m - 1) }
    else{ pch <- rep(pch, length.out = m) }
    if (is.null(col)) { col <- 1:m }
    else{ col <- rep(col, length.out = m) }

    if (is.null(legendtext) & !is.null(axistext)) \label{def3}
        legendtext <- axistext
    if (!is.null(legendtext) & is.null(axistext))
        axistext <- legendtext
    if (is.null(legendtext) & is.null(axistext) & !is.null(names(xp)))
        legendtext <- names(xp)
    if (is.null(legendtext) & is.null(axistext) & !is.null(names(coefs)))
        legendtext <- names(coefs)
    if (is.null(axistext) & !is.null(colnames(x)))
        axistext <- colnames(x)
    if (is.null(axistext))
        axistext <- legendtext

    ones <- matrix(1, nrow = n, ncol = 1)
    y <- xp - ones %*% refvalues

    if (!is.null(confinterval)){
        upp_ci <- confinterval$upp - ones %*% confinterval$upprefvalues
```

```
        upp_ci <- upp_ci[select]
        low_ci <- confinterval$low - ones %*% confinterval$lowrefvalues
        low_ci <- low_ci[select]
    }
    if (identical(plottype, "hazard")){ \label{def4}
        y <- exp(y)
    }
    if (identical(plottype, "hazard") & !is.null(confinterval)){
        low_ci <- exp(low_ci)
        upp_ci <- exp(upp_ci)
    }
    yrange <- y    # makes sure that confidence intervals fit to the screen
    if (!is.null(confinterval)){
        yrange <- as.data.frame(c(y, low_ci, upp_ci))
    }

    if (length(ylim) != 2){
        maxy <- max(yrange, na.rm = TRUE)
        miny <- min(yrange, na.rm = TRUE)
    }else{
        maxy <- ylim[2]
        miny <- ylim[1]
    }

    if (identical(plottype, "hazard")) {
        if (is.null(ylab)) ylab <- "relative hazard"
        if (is.null(yticks))
```

```
            yticks <- c(pretty(c(miny, 1)), pretty(c(1, maxy)))
        reftickvalue <- 1
        logvar = "y"
    }
    if (identical(plottype, "loghazard")) { \label{def5}
        if (is.null(ylab)) ylab <- "logarithm of the relative hazard"
        if (is.null(yticks))
            yticks <- c(pretty(c(miny, 0)), pretty(c(0, maxy)))
        reftickvalue <- 0
        logvar = ""
    }

    if (is.null(yvalues)) yvalues <- yticks
    quantiles <- c(0, 0.25, 0.5, 0.75, 1)

### Output to console ####
    A <-matrix(0, m, 5)
    colnames(A) <- c("Min.", "1st Qu.", "Median" , "3rd Qu.", "Max.")
    rownames(A) <- legendtext

    ind <- NULL
    for (i in 1:m){
        ordered <- order(x[, i], na.last = TRUE)
        ind <- cbind(ind, ordered)   # ind is used later
        A[i,] <- quantile(y[, i], probs = quantiles, na.rm = TRUE)
    }
    cat("Y-axis range: ", signif(c(miny, maxy), 3), "\n", "\n")
```

```
    if (identical(plottype, "hazard")) cat("Relative hazards for each covarite:", "\n")
    if (identical(plottype, "loghazard")) cat("Logarithm of the relative hazards for each covarite:", "\n")
    print(signif(A, 3))
###
    nasum <- colSums(is.na(x)) \label{def6}

    for (j in 1:m) {
        nj <- n - nasum[j]
        ranks <- seq(0, 1, length = nj)     # scales the values to [0,1]
        places <- quantile(1:nj, probs = quantiles)    # quantiles of ranks

        if (j == 1) {
            plot(1, 1, ylim = c(miny, maxy), xlim = c(0, 1),
              xlab = "", ylab = ylab, type = "n", col = col[j],
              pch = pch[j], lwd = lwd[j], lty = lty[j], axes = FALSE, log = logvar,
              ...)

            axis(1, at = quantiles, labels = FALSE)    # marks ticks on x-axis
            axis(2, at = yticks, labels = FALSE)    # marks ticks on y-axis
            axis(2, at = yvalues, labels = as.character(yvalues))    # marks values on y-axis
            box()

            if (reftick)    # eboldens the reference tick
                axis(2, at = reftickvalue, labels = FALSE, lwd.ticks = 2)

            if (refline)    # draws the reference line
                abline(h = reftickvalue,  col = refline.col, lty = refline.lty, lwd = refline.lwd)
```

```
        }



        lines(ranks, y[ind[1:nj, j], j], col = col[j], lwd = lwd[j], lty = lty[j], ...) \label{def7}
        points(quantiles, y[ind[1:nj, j], j][places], col = col[j], pch = pch[j], cex = cex[j], bg = bg[j],
            lwd = pt.lwd[j], ...)

        xlabels <- x[ind[places, j], j]    # quantiles for covariate j
        if (is.numeric(xlabels)) xlabels <- signif(xlabels, 3)

        mtext(side = 1, at = c(axistextposition, quantiles),
          adj = c(1,rep(0.5, length(quantiles))), text = c(axistext[j], as.character(xlabels)), line = j)

        if (!is.null(confinterval)) {
            lines(ranks, low_ci[ind[1:nj, j], j], col = col[j], lwd = lwd[j], lty = lty[j] + 1, ...)
            points(quantiles, low_ci[ind[1:nj, j], j][places], col = col[j], pch = pch[j], cex = cex[j], bg = bg[j],
                lwd = pt.lwd[j],...)
            lines(ranks, upp_ci[ind[1:nj, j], j], col = col[j], lwd = lwd[j], lty = lty[j] + 1,...)
            points(quantiles, upp_ci[ind[1:nj, j], j][places], col = col[j], pch = pch[j], cex = cex[j], bg = bg[j],
                lwd = pt.lwd[j],...)
        }
    }

    if (!is.null(confinterval)){
### Output to console ####
        cat("\n")
```

```
        if (identical(plottype, "hazard"))
            cat("Relative hazards for the confidence intervals of each covariate:", "\n")
        if (identical(plottype, "loghazard"))
            cat("Logarithm of the relative hazards for the confidence intervals of each covariate:", "\n")
        B <- matrix(0, 2 * m, 5)  \label{def8}
        colnames(B) <- c("Min.", "1st Qu.", "Median" , "3rd Qu.", "Max.")
        low_legend <- paste("Low", legendtext, sep = "_")
        upp_legend <- paste("Upp", legendtext, sep = "_")
        rownames(B)[2 * 1:m] <- upp_legend
        rownames(B)[2 * 1:m - 1] <- low_legend

        for (i in 1:m)
            B[2 * i - 1,] <- quantile(low_ci[, i], probs = quantiles, na.rm = TRUE)

        for (i in 1:m)
            B[2 * i,] <- quantile(upp_ci[, i], probs = quantiles, na.rm = TRUE)

        print(signif(B, 3))
###############
    }

    legend(legendlocation, legend = legendtext, col = col, lwd = lwd,
      pch = pch, lty = lty, bty = "n", pt.cex = cex, pt.lwd = pt.lwd, pt.bg = bg)

}
```

D.3 rankhazardplot.coxph

```
rankhazardplot.coxph <- function (
    coxphobj, data = NULL, select = NULL, refpoints = NULL,
    CI_level = 0.95, x_CI = NULL, confint = FALSE, legendtext = NULL,
    axistext = NULL, legendlocation = "top", axistextposition = -0.1,
    reftick = TRUE, refline = FALSE, refline.col = 1, refline.lwd = 1,
    refline.lty = 2, ylab = NULL, ylim = NULL, yticks = NULL,
    yvalues = NULL, plottype = "hazard", na.rm = TRUE, draw = TRUE,
    return = FALSE, col = NULL, lwd = 1, lty = 1, pch = NULL,
    cex = 1, bg = "transparent", pt.lwd = 1, ...)
{
    if (is.null(data))
        stop("Covariate data need to be provided as argument data.")

    if (is.null(coxphobj$x) & is.null(x_CI))
        stop("To calculate confidence intevals covariate data need to be provided either as argument x_CI
         or as coxphobj$x.")

    if (is.null(x_CI))
        x_CI <- as.data.frame(coxphobj$x)

    term_labels <- attr(coxphobj$terms, "term.labels")

    if (is.null(select))
        select <- 1:length(term_labels)

    trans_var <- which(!is.element(term_labels, names(data))) # checks which labels are different as in the data
```

```
    # changes the labels of the transformed variables to get the same labels as in the data
    labels <- sapply(strsplit(term_labels[trans_var], ',', fixed = TRUE), '[[', 1) # extracts the string before ','
    labels <- gsub("^.*\\(", "", labels)  # deletes all characters before '('
    labels <- gsub(")", "", labels, fixed = TRUE)  # deletes all ')'
    # combines original and changed labels
    data_labels <- term_labels
    data_labels[trans_var] <- labels
    x <- data[data_labels]

    if (na.rm) x <- na.omit(x)

    factorlevels <- coxphobj$xlevels
    factorlabs <- names(factorlevels)
    factors <- which(is.element(term_labels, factorlabs))
    nonfactors <- which(!is.element(term_labels, factorlabs))

    refs <- data.frame(setNames(replicate(length(data_labels), numeric(0), simplify = F), data_labels))

    for (i in nonfactors)
        refs[1, i] <- median(x[, i], na.rm = TRUE)

    j <- 1
    for (i in factors){
        refs[1, i] <- factorlevels[[j]][1]
        refs[i] <- factor(refs[i], levels = factorlevels[[j]])
        j <- j + 1
    }
```

```
    if(!is.null(refpoints)){
        change <- which(!is.na(refpoints))
        if(is.numeric(refpoints)){
            refs[1, select[change]] <- refpoints[change]
        }else{
            refs[1, intersect(factors, select[change])]
                <- refpoints[is.element(select[change], intersect(factors, select[change]))]
            refs[1, intersect(nonfactors, select[change])]
                <- as.numeric(refpoints[is.element(select[change], intersect(nonfactors, select[change]))])
        }
    }

    newdata <- x
    predictions <- predict(coxphobj, type = "terms", newdata = newdata)
    newdata <- refs
    pred_refvalues <- predict(coxphobj, type = "terms", newdata = newdata)

    refvalues <- as.vector(pred_refvalues)
    names(refvalues) <- attr(pred_refvalues, "dimnames")[[2]]

    xp <- as.data.frame(predictions)

### Calculating the confidence intervals ###

    coefslow <- confint(coxphobj, level = CI_level)[, 1]
    coefsupp <- confint(coxphobj, level = CI_level)[, 2]

```

```
    refs[factors] <- 0
    refs <- as.vector(as.matrix(refs))

    Values <- coxph_CI(coxphobj, x_CI, coxphobj$coef, refs)
    CIlow <- coxph_CI(coxphobj, x_CI,  coefslow, refs)
    CIupp <- coxph_CI(coxphobj, x_CI, coefsupp, refs)

    confinterval <- list(x = Values$x, xp = Values$xp, refvalues = Values$refvalues, low = CIlow$xp,
                         lowrefvalues = CIlow$refvalues, upp = CIupp$xp, upprefvalues = CIupp$refvalues)
    select_CI <- Values$select_CI
    selecttext <- select

    if (confint){
        CI <- confinterval
        select <- which(is.element(select_CI, select))
        selecttext <- select_CI[select]
        if(length(select) == 0)
            stop("Confidence intevals cannot be caluclated for selected covariates.")
    } else {
        CI<- NULL
    }
    if (!is.null(legendtext) & is.null(axistext)){
        axistext <- legendtext
    }
    if (is.null(legendtext) & !is.null(axistext)){
        legendtext <- axistext
    }
```

```
    if (is.null(legendtext)){
        legendtext <- term_labels[selecttext]
        axistext <- data_labels[selecttext]
    }

    if (draw)
        rankhazardplot.default(
            x = x[select], xp = xp[select], refvalues = refvalues[select],
            legendtext = legendtext, axistext = axistext,
            na.rm = na.rm, select = select, confinterval = CI,
            legendlocation = legendlocation, axistextposition = axistextposition,
            reftick = reftick, refline = refline, refline.col = refline.col,
            refline.lwd = refline.lwd, refline.lty = refline.lty, ylab = ylab,
            ylim = ylim, yticks = yticks, yvalues = yvalues, plottype = plottype,
            col = col, lwd = lwd, lty = lty, pch = pch,
            cex = cex, bg = bg, pt.lwd = pt.lwd, ...)

    if (return)
        return(list(x = x, xp = xp, refvalues = refvalues, confinterval = confinterval))
}
```

D.4 rankhazardplot.cph

```
rankhazardplot.cph <- function (
    cphobj, data = NULL, select = NULL, refpoints = NULL,
    CI_level = 0.95, x_CI = NULL, confint = FALSE, legendtext = NULL,
    axistext = NULL, legendlocation = "top", axistextposition = -0.1,
    reftick = TRUE, refline = FALSE, refline.col = 1, refline.lwd = 1,
    refline.lty = 2, ylab = NULL, ylim = NULL, yticks = NULL,
    yvalues = NULL, plottype = "hazard", na.rm = TRUE, draw = TRUE,
    return = FALSE, col = NULL, lwd = 1, lty = 1, pch = NULL,
    cex = 1, bg = "transparent", pt.lwd = 1, ...)
{
    if (is.null(data))
        stop("Covariate data need to be provided as argument data.")

    if (is.null(cphobj$x) & is.null(x_CI))
        stop("To calculate confidence intevals covariate data need to be provided either as argument x_CI
           or as cphobj$x.")

    if (is.null(x_CI))
        x_CI <- as.data.frame(cphobj$x)

    term_labels <- cphobj$Design$name

    if (is.null(select))
        select <- 1:length(term_labels)

    data_labels <- term_labels
```

```
    x <- data[data_labels]

    if (na.rm) x <- na.omit(x)

    factorlevels <- cphobj$Design$parms
    factorlabs <- names(factorlevels)
    factors <- which(is.element(term_labels, factorlabs))
    nonfactors <- which(!is.element(term_labels, factorlabs))

    refs <- data.frame(setNames(replicate(length(data_labels), numeric(0), simplify = F), data_labels))

    for (i in nonfactors)
     refs[1, i] <- median(x[, i], na.rm = TRUE)

    j <- 1
    for (i in factors){
        refs[1, i] <- factorlevels[[j]][1]
        refs[i] <- factor(refs[i], levels  = factorlevels[[j]])
        j <- j + 1
    }

    if (!is.null(refpoints)){
        change <- which(!is.na(refpoints))
        if (is.numeric(refpoints)){
            refs[1, select[change]] <- refpoints[change]
        }else{
            refs[1, intersect(factors, select[change])]
```

```
                <- refpoints[is.element(select[change], intersect(factors, select[change]))]
            refs[1, intersect(nonfactors, select[change])]
                <- as.numeric(refpoints[is.element(select[change], intersect(nonfactors, select[change]))])
        }
    }

    predictions <- predict(cphobj, type = "terms", newdata = x)
    pred_refvalues <- predict(cphobj, type = "terms", newdata = refs)

    refvalues <- as.vector(pred_refvalues)
    names(refvalues) <- attr(pred_refvalues, "dimnames")[[2]]

    xp <- as.data.frame(predictions)

### Calculating the confidence intervals ###

    coefslow <- confint(cphobj, level = CI_level)[, 1]
    coefsupp <- confint(cphobj, level = CI_level)[, 2]

    refs[factors] <- 0
    refs <- as.vector(as.matrix(refs))

    Values <- cph_CI(cphobj, x_CI, cphobj$coef, refs)
    CIlow <- cph_CI(cphobj, x_CI, coefslow, refs)
    CIupp <- cph_CI(cphobj, x_CI, coefsupp, refs)
```

```
    confinterval <- list(x = Values$x, xp = Values$xp, refvalues = Values$refvalues, low = CIlow$xp,
        lowrefvalues = CIlow$refvalues, upp = CIupp$xp, upprefvalues = CIupp$refvalues)
    select_CI <- Values$select_CI
    selecttext <- select

    if (confint){
        CI <- confinterval
        select <- which(is.element(select_CI, select))
        selecttext <- select_CI[select]
        if(length(select) == 0)
            stop("Confidence intevals cannot be caluclated for selected covariates.")
    } else {
        CI <- NULL
    }

    if (!is.null(legendtext) & is.null(axistext)){
        axistext <- legendtext
    }
    if (is.null(legendtext) & !is.null(axistext)){
        legendtext <- axistext
    }

    if (is.null(legendtext)){
        legendtext <- attr(cphobj$terms, "term.labels")[selecttext]
        axistext <- term_labels[selecttext]
    }
```

```
    if (draw)
        rankhazardplot.default(
            x = x[select], xp = xp[select], refvalues = refvalues[select],
            legendtext = legendtext, axistext = axistext,
            na.rm = na.rm, select = select, confinterval = CI,
            legendlocation = legendlocation, axistextposition = axistextposition,
            reftick = reftick, refline = refline, refline.col = refline.col,
            refline.lwd = refline.lwd, refline.lty = refline.lty, ylab = ylab,
            ylim = ylim, yticks = yticks, yvalues = yvalues, plottype = plottype,
            col = col, lwd = lwd, lty = lty, pch = pch,
            cex = cex, bg = bg, pt.lwd = pt.lwd, ...)

    if (return)
        return(list(x = x, xp = xp, refvalues = refvalues, confinterval = confinterval))
}
```

D.5 coxph_CI

```
457 coxph_CI <- function(coxphobj, x, coefs, refpoints){
458
459 # This function calculates predictions as a product of coefs and x, and
460 # reference values as a product of coefs and refvalues.
461 # Factors are given in a dummy format but returned as one variable.
462 # The function returns data as x, predictions as xp, reference values as refvalues
463 # and indices of covariates for which the predictions can be calculated by this function as
464 # select_CI.
465
466     factorlevels <- coxphobj$xlevels
467     factorlabs <- names(factorlevels)
468
469     xp <- as.data.frame(t(coefs * t(x)))
470
471     covariatelabs <- attr(coxphobj$terms, "term.labels")
472     factors <- which(is.element(covariatelabs, factorlabs))
473
474     columns <- sapply(coxphobj$assign, length)
475     orig_var <- which(columns == 1)
476
477     select <- sort(union(orig_var, factors))        #for these covariates confidence intervals can be calculated
478     indices <- sapply(coxphobj$assign, "[[", 1)        #indices where values for each covariate are/begin
479
480     xp[factorlabs] <- 0        #these covariates don't exist yet because factors are in a dummy format
481     x[factorlabs] <- 0
482
```

```
    j <- 1
    for (i in factors){    # levels of factors are combined to one variable

        if (columns[i] > 1)
            xp[names(columns)[i]] <- apply(xp[, indices[i] + 0:(columns[i] - 1)], 1, sum)
        else xp[names(columns)[i]] <- xp[, indices[i]]

        for (l in 1:(columns[i])){   #the values of different levels are copied into same variable
            x[factorlabs[j]] <- ifelse(x[, indices[i] + l - 1] == 1, factorlevels[[j]][l + 1], x[, factorlabs[j]])
        }
        #the cases with xp == 0 have the value of the reference level
        x[factorlabs[j]] <- ifelse(xp[, factorlabs[j]] == 0, factorlevels[[j]][1], x[, factorlabs[j]])
        j <- j + 1
    }

    #the factors in the model are coerced as factors in x

    j <- 1
    for (i in factors) {
        xfactor <- as.factor(x[, factorlabs[j]])
        x[factorlabs[j]] <- relevel(xfactor, ref = factorlevels[[j]][1])
        j <- j + 1
    }
```

```
    covariatelabs <- covariatelabs[select]

    xp <- xp[covariatelabs]
    x <- x[covariatelabs]

    refvalues <- coefs[indices[select]] * refpoints[select]

    names(refvalues) <- covariatelabs
    refvalues[factorlabs] <- 0          # the reference value for factors is zero as coefficient for the reference
                                          # level is zero

    return(list(x = x, xp = xp, refvalues = refvalues, select_CI = select))
}
```

D.6 cph_CI

```
cph_CI <- function(cphobj, x, coefs, refpoints){

# This function calculates predictions as a product of coefs and x, and
# reference values as a product of coefs and refvalues.
# Factors are given in a dummy format but returned as one variable.
# The function returns data as x, predictions as xp, reference values as refvalues
# and indices of covariates for which the predictions can be calculated by this function as
# select_CI.

    factorlevels <- cphobj$Design$parms
    factorlabs <- names(factorlevels)

    xp <- as.data.frame(t(coefs * t(x)))

    covariatelabs <- cphobj$Design$name
    factors <- which(is.element(covariatelabs, factorlabs))

    columns <- sapply(cphobj$assign, length)
    orig_var <- which(columns == 1)

    select <- sort(union(orig_var, factors))        #for these covariates confidence intervals can be calculated
    indices <- sapply(cphobj$assign, "[[", 1)        #indices where values for each covariate are/begin

    xp[factorlabs] <- 0        #these covariates don't exist yet because factors are in a dummy format
    x[factorlabs] <- 0

```

```
    j <- 1
    for (i in factors){  # levels of factors are combined to one variable

        if (columns[i] > 1)
            xp[names(columns)[i]] <- apply(xp[, indices[i] + 0:(columns[i] - 1)], 1, sum)
        else xp[names(columns)[i]] <- xp[, indices[i]]

        for (l in 1:columns[i]){   #the values of different levels are copied into same variable
            x[factorlabs[j]] <- ifelse(x[, indices[i] + l - 1] == 1, factorlevels[[j]][l + 1], x[, factorlabs[j]])
        }
        #the cases with xp == 0 have the value of the reference level
        x[factorlabs[j]] <- ifelse(xp[, factorlabs[j]] == 0, factorlevels[[j]][1], x[, factorlabs[j]])
        j <- j + 1
    }

    #the factors in the model are coerced as factors in x

    j <- 1
    for (i in factors) {
        xfactor <- as.factor(x[, factorlabs[j]])
        x[factorlabs[j]] <- relevel(xfactor, ref = factorlevels[[j]][1])
        j <- j + 1
    }
```

```
    covariatelabs <- covariatelabs[select]

    xp <- xp[covariatelabs]
    x <- x[covariatelabs]

    refvalues <- coefs[indices[select]] * refpoints[select]

    names(refvalues) <- covariatelabs
    refvalues[factorlabs] <- 0 # the reference value for factors is zero as coefficient for the reference
                                  # level is zero

    return(list(x = x, xp = xp, refvalues = refvalues, select_CI = select))
}
```

E Start–stop-aineiston luominen

```
time <- gero$eloaika
delta <- gero$delta
# uusien mittauksien voimaantuloaika
365*5+1 #1826

time1 <- NULL
time2 <- NULL
status <- NULL

sukup <- NULL
vaik <- NULL
nopeus <- NULL
bmi <- NULL

indeksi <- NULL
mittauskerta <- NULL


for(i in 1:dim(gero)[1]){ #jokainen henkilö käydään läpi

    # ne missä time1 == 0, alkaa uusi hlö
    # asetetaan ensimmäiseksi aikahavainnoksi 0
    time1 <- c(time1, 0)

    if(time[i] < 1826)
        time2 <- c(time2, time[i])
    else{
        time1 <- c(time1, 1826)
        time2 <- c(time2, 1826, time[i])
    }

    # hlö kuolee, jos status = 1, sensuroituu, jos 0
    # ainoat sensuroinnit on tutkimusajan lopussa
    if(time[i] < 1826)
        status <- c(status, 1)
    else
        if(time[i] >= 1826 & delta[i] == 1)
            status <- c(status, c(0,1))
        else
            status <- c(status, c(0,0))

    bmi <- c(bmi, gero$bmi75[i])
```

```
42      vaik <- c(vaik, gero$vaik75[i])
43      nopeus <- c(nopeus, gero$nopeus75[i])
44      sukup <- c(sukup, as.character(gero$sp[i]))
45      indeksi <- c(indeksi, i)
46      mittauskerta <- c(mittauskerta, 1)
47
48      if(time[i] >= 1826){
49          bmi <- c(bmi, gero$bmi80[i])
50          vaik <- c(vaik, gero$vaik80[i])
51          nopeus <- c(nopeus, gero$nopeus80[i])
52          sukup <- c(sukup, as.character(gero$sp[i]))
53
54          indeksi <- c(indeksi, i)
55          mittauskerta <- c(mittauskerta, 2)
56      }
57  }
58
59  gero_aika <- data.frame(time1, time2, indeksi,status,  mittauskerta,
60  sukup, bmi, vaik, nopeus)
```

Lähteet

Allison, Paul D. (1995). *Survival Analysis Using SAS®: A Practical Guide*. SAS Institute Inc.: Cary, NC.

Barlow, William E. & Prentice, Ross L. (1988). Residuals for Relative Risk Regression. *Biometrika*, **75**, 65–74.

Cain, Kevin C. & Lange, Nicholas T. (1984). Approximate Case Influence for the Proportional Hazards Regression Model with Censored Data. *Biometrics*, **40**, 493–499.

Cleves, Mario; Gould, William; Gutierrez, Roberto & Marchenko, Yulia (2008). *An Introduction to Survival Analysis Using Stata*. 2. painos. Stata Press: Texas.

Clinical Reference Laboratory, Inc. (2014). `http://www.crlcorp.com/test/total-bilirubin/`. Viitattu 14.9.2014.

Collett, David (2003). *Modelling Survival Data in Medical Research*. 2. painos. Chapman & Hall/CRC: Lontoo.

Cox, David R. (1972). Regression Models and Life-Tables (with discussion). *Journal of the Royal Statistical Society*, B, **34**, 187–220.

Efron, Bradley (1977). The Efficiency of Cox's Likelihood Function for Censored Data. *Journal of the American Statistical Association*, **72**, 557–565.

Eilers, Paul H. & Marx, Brian D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, **11**, 89-121.

Fleming, Thomas R. & Harrington, David P. (1991). *Counting processes and survival analysis*. Wiley: New York.

Färkkilä, Martti (1993). Primaarinen sklerosoiva kolangiitti. *Lääketieteellinen Aikakauskirja Duodecim*, **4**.

Gerontologian tutkimuskeskus. Evergreen-project. `http://www.gerec.fi/en/research/health-functioning-and-longevity/evergreen-project`. Viitattu 2.1.2015.

Gerontologian tutkimuskeskus. Ikivihreät-projekti. `http://www.gerec.fi/tutkimus/terveys-toimintakyky/evergreen-project`. Viitattu 2.1.2015.

Grambsch, Patricia M. & Therneau, Terry M. (1994). Proportional Hazards Tests and Diagnostics Based on Weighted Residuals. *Biometrika*, **81**, 515–526.

Harrell Jr., Frank E. (2014). rms: Regression Modeling Strategies. R package version 4.2-0. `http://CRAN.R-project.org/package=rms`

Karvanen, Juha & Harrell Jr., Frank E. (2009). Visualizing covariates in proportional hazards model. *Statistics in Medicine*, **28**, 1957–1966.

Karvanen, Juha (2012). rankhazard: Rank-hazard plots. R package version 0.8-1. `http://cran.r-project.org/src/contrib/Archive/rankhazard/`

Karvanen, Juha & Koski, Nanni (2014). rankhazard: Rank-hazard plots. R package version 1.0. `http://CRAN.R-project.org/package=rankhazard`

Lee, Elisa T. & Wang, John W. (2003). *Statistical Methods for Survival Data Analysis*. 3. painos. Wiley: New Jersey.

Locke III, G. Richard; Therneau, Terry M.; Ludwig Jurgen; Dickson, E. Rolland; & Lindor, Keith D. (1996). Time Course of Histological Progression in Primary Biliary Cirrhosis. *Hepatology*, **23**, 52–56.

Läärä, Esa; Luostarinen, Tapio; Hakulinen, Timo; Lyytikäinen, Outi; Sarna, Seppo; Virtala, Anna-Maija; Riihimäki, Hilkka & Hakama Matti, Epidemiologian englanti-suomi-englanti-sanasto. `http://www.finepi.org/files/englantisuomi.pdf`, Suomen Epidemiologian Seura ja Duodecim 26.11.2008. Viitattu 24.11.2014.

Mustajoki, Pertti. Sappikirroosi. `http://www.terveyskirjasto.fi`. Lääkärikirja Duodecim. Duodecim 9.12.2013. Viitattu 10.8.2014.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. `http://www.R-project.org/`.

Reinikainen, Jaakko; Laatikainen Tiina; Karvanen Juha & Tolonen Hanna (2014). Lifetime cumulative risk factors predict cardiovascular disease mortality in a 50-year follow-up study in Finland. *International Journal of Epidemiology*. Esijulkaisu Internetissä, DOI: `10.1093/ije/dyu235`.

Schoenfeld, David (1982). Partial Residuals for The Proportional Hazards Regression Model. *Biometrika*, **69**, 239–241.

Therneau, Terry (2014). A Package for Survival Analysis in S. R package version 2.37-7. `http://CRAN.R-project.org/package=survival`.

Therneau, Terry M. & Grambsch, Patricia M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer: New York.

Therneau, Terry M.; Grambsch, Patricia M. & Fleming, Thomas R. (1990). Martingale based residuals for survival models. *Biometrika*, **77**, 147–160.

Tufte, Edward R. (1983). *The Visual Display of Quantitative Information*. Graphics Press: Connecticut.