SYNTYMÄKOHORTTI 1966 OARSI-NIVELRIKKOLUOKITUKSEN TOISTETTAVUUS

Yliklaavu, Seppo
Syventävien opintojen tutkielma
Lääketieteellinen tiedekunta
Oulun yliopisto
Marraskuu 2015
Prof. Jaakko Niinimäki

OULUN YLIOPISTO
Lääketieteellinen tiedekunta
Lääketieteen koulutusohjelma

TIIVISTELMÄ

Yliklaavu, Seppo: Syntymäkohortti 1966 OARSI-nivelrikkoluokituksen toistettavuus
Syventävien opintojen tutkielma: 20 sivua

Nivelrikko on merkittävä ja yleisin niveliin kohdistuva sairaus, jonka tarkkaa etiologiaa ei kuitenkaan tunneta. Nivelrikon aiheuttamia tyypillisiä muutoksia ovat nivelruston rappeutuminen, nivelvälin kaventuminen, nivelen kipeytyminen ja potilaan toimintakyvyn lasku. Nivelrikon diagnostiikassa natiiviröntgenkuvilla on keskeinen asema ja nivelrikon aiheuttamille muutoksille on kehitetty erilaisia radiologisia luokitusmenetelmiä.

Tutkimuksen tarkoituksena oli arvioida OARSI-luokitusjärjestelmän kohdalla tutkijoiden välistä että sisäistä toistettavuutta keskittyen polvinivelen muutosten luokitteluun. Tutkimuksen aineistona toimi osaotos Pohjois-Suomen syntymäkohortti 1966:sta. Tutkijoina toimi kaksi lääketieteen opiskelijaa, jotka luokittelivat noin 1900 polvea arvioiden osteofyyttimuodostusta sekä nivelraon kaventumaa OARSI-järjestelmään perustuen. Luokittelutuloksista laskettiin erilaisia toistettavuuden tilastollisia mittareita, joista yleisesti käytetyimpiä on kappa-arvo.

Tutkijoiden välinen toistettavuus vaihteli huonosta kohtalaiseen riippuen mitattiinko yksittäisiä luokiteltavia tekijöitä, OARSI-kriteereitä vai lopullista diagnoosia ja saadut kappa-arvot vaihtelivat välillä 0.084–0.61. Tutkijoiden sisäinen toistettavuus oli parempi ja saadut kappa-arvot vaihtelivat tutkijakohtaisesti välillä 0.25–1.0 sekä 0.0–1.0 ollen jälkimmäisellä tutkijalla kuitenkin keskimäärin korkeampia. Tutkijoiden tavassa käyttää eri luokitusarvoja havaittiin myös eroavaisuutta erityisesti osteofyyttiluokituksen tapauksessa.

Tutkimuksessa saadut tulokset vahvistavat jo aikaisemmissa tutkimuksissa osoitettua, että radiologisen luokitusjärjestelmän toistettavuuden arviointi on riippuvaista käytetystä luokitusjärjestelmästä, käytetystä aineistosta sekä itse tutkijoista. Siten eri tutkimusten tuloksiin täytyy suhtautua varauksella ottaen huomioon edellä mainitut tekijät. Joissakin tapauksissa on syytä myös ymmärtää käytettyjen tilastollisten mittareiden kuten kappa-arvon käyttäytymisen riippuvuus muuttujien jakautumisesta.

Avainsanat: polvi, nivelrikko, OARSI, luokitus, toistettavuus, kappa

SISÄLLYS

1. JOHDANTO

Nivelrikko eli artroosi on yleisin niveliin kohdistuva sairaus ja sen kansantaloudellinen kustannusvaikutus on merkittävä (Heliövaara ym. 2008). Nivelrikko kohdistuu usein polveen ja sille tyypillisiä muutoksia ovat nivelruston rappeutuminen, nivelvälin kaventuminen, nivelen kipeytyminen ja potilaan toimintakyvyn lasku. Polven nivelrikkoa ei juurikaan tavata alle 40-vuotialla ja sen esiintyvyys lisääntyy iän myötä. Varsin usein se voi kuitenkin olla oireeton. Nivelrikon syitä ei tunneta tarkasti, mutta mekaanisilla tekijöillä on keskeinen merkitys (Jurvelin ym. 2008). Joka tapauksessa rustoa hajottavat biokemialliset tekijät saavat yliotteen korjaavista prosesseista.

Polvinivelen natiiviröntgenkuvaus on hyvin tyypillinen menetelmä nivelrikon diagnostiikassa (Jurvelin ym. 2008). Sen avulla voidaan selvittää nivelraon kapenemista, rustonalaisen luun skleroosia, osteofyyttimuodostuksen kaltaisia rappeumamuutoksia, nivelen virheasentoa sekä irtokappaleiden ilmenemistä nivelessä. Näkyäkseen röntgenkuvassa nivelrikkomuutosten täytyy kuitenkin olla jo kohtalaisen pitkällä.

Nivelrikon ensisijainen hoito on aina lääkkeetöntä ja konservatiivista. Kirurgiseen hoitoon tulisi suhtautua harkiten ja ainoastaan vaikeimmissa tapauksissa (Konttinen ym. 2003, Remes ym. 2008). Ensisijaisiin hoitokeinoihin kuuluvat siten esimerkiksi laihdutus, lihasharjoitteet, fysioterapia ja erilaisten apuvälineiden käyttäminen. Lääkkeellisen hoidon piiriin kuuluvat erilaiset kipulääkkeet, tulehduskipulääkkeet sekä paikallisesti niveleen injektoitavat glukokortikoidit. Kirurgisesta hoidoista mainittakoon artroskopia, osteotomia sekä osa- ja kokotekonivelleikkaukset (Remes ym. 2008). Artroskopiassa tehdään tarkka nivelen sisäinen tutkimus sekä muovataan kuluneita rakenteita. Osteotomiassa kuormitus pyritään siirtämään vaurioituneelta nivelalueelta terveelle pinnalle esimerkiksi katkaisemalla tibia ja kääntämällä se valgukseen. Tulevaisuudessa voidaan mahdollisesti korvata vaurioitunutta ja heikosti uusiutuvaa rustokudosta täysin uudella rustokudoksella.

Nivelrikkoa varten on olemassa useita radiologisia luokitusmenetelmiä, joista jo varsin iäkäs Kellgren & Lawrence -luokitus on edelleen yleisimpiä ja sen toistettavuuden arvioidaan olevan kohtalainen (Kellgren Jh ja Lawrence Js 1957, Ojala ja Arokoski 2012). Kellgren & Lawrence -luokitusta kohtaan on kuitenkin esitetty kritiikkiä ja vaihtoehtoisia luokitusmenetelmiä on kuvattu kirjallisuudessa sen tilalle (Altman RD ja Gold GE 2007, Culvenor ym. 2014). Tässä

tutkimuksessa on tarkoitus arvioida Kellgren & Lawrence -järjestelmälle vaihtoehtoisen luokitusjärjestelmän toistettavuutta syntymäkohortti 1966 perustuvan aineiston pohjalta.

Käytettäessä radiologisia luokitusmenetelmiä nivelrikon diagnostiikassa, niiden tarjoama informaatio yhdistetään potilaan muuhun kliiniseen kuvaan. Diagnostiikan lisäksi luokitusmenetelmiä voidaan käyttää nivelrikon progression seurannassa, erilaisten hoitomenetelmien vaikuttavuuden arvioinnissa sekä erilaisissa nivelrikon riskitekijöitä kartoittavissa tutkimuksissa.

2. TUTKIMUKSEN TARKOITUS JA TUTKIMUSONGELMAT

Tutkimuksen tarkoituksena oli arvioida radiologisin perustein tapahtuvassa polven nivelrikkoluokittelussa luokittelijoiden välistä ja sisäistä toistettavuutta sekä luotettavuutta erilaisin mittarein. Tässä tutkimuksessa käytettävä luokittelumenetelmä on Osteoarthritis Research Society Internationalin (OARSI) julkaisema ja se on kuvattu kirjallisuudessa atlasmaisesti (Altman RD ja Gold GE 2007). Tällaisen luokittelun tulos ja mahdollinen nivelrikkodiagnoosi on siis riippuvainen käytetystä luokittelujärjestelmästä, mutta erityisesti myös henkilöistä tai tutkijoista, jotka tekevät varsinaisen luokittelun.

Polvinivelen OARSI-luokitukseen liittyvät tekijät, jotka vaikuttavat lopulliseen nivelrikkodiagnoosiin, ovat osteofyyttimuodostus sekä nivelraon kaventuma. Nämä tekijät arvioidaan sekä mediaaliselle että lateraaliselle tibiofemoraaliselle kompartmentille. Lopullinen diagnoosi muodostuu näistä osatekijöistä muodostetun kolmen eri kriteerin perusteella.

Taulukko 1 esittää, kuinka polvinivelen osteofyyttimuodostus sekä nivelraon kaventuminen arvioidaan OARSI-luokituksessa neliportaisella asteikolla. Molemmat arvioidaan erikseen femurille että tibialle sekä mediaalisesti että lateraalisesti. Luokitusarvot muodostavat ordinaaliasteikon, jossa muuttuja voidaan järjestää luokitusarvon perusteella, mutta luokkien välistä eroa ei voida yksiselitteisesti mitata kuten intervalliasteikollisella muuttujalla (esimerkiksi lämpötila).

Taulukko 1. OARSI-luokituksessa käytettävät luokat osteofyyteille sekä nivelraon kaventumalle (Altman RD ja Gold GE 2007, Culvenor ym. 2014).

Luokka	Kuvaus
0	ei lainkaan
1	lievä
2	kohtalainen
3	vakava

Lopullinen polvinivelen nivelrikkodiagnoosi riippuu kolmesta eri kriteeristä (Taulukko 2). Mikäli yksikin näistä kriteereistä täyttyy mediaalisesti tai lateraalisesti, katsotaan nivelessä olevan nivelrikko eli diagnoosi on positiivinen. Kriteerien ja diagnoosin voidaan katsoa muodostavan yksinkertaisimman ordinaalisen asteikon muuttujan ollessa dikotominen.

Taulukko 2. Alkuperäiset OARSI-kriteerit, jotka muodostetaan osteofyytti- ja nivelrakoluokitusten perusteella (Culvenor ym. 2014).

OARSI-kriteeri no.	Kuvaus
1	Nivelraon kaventuman luokka ≥ 2
2	Osteofyyttiluokkien summa ≥ 2
3	Luokan 1 nivelraon kaventuma yhdessä luokan 1 osteofyytin kanssa

3. TUTKIMUSAINEISTO JA TUTKIMUSMENETELMÄT

3.1. Aineisto ja tutkimuksen kulku

Tutkimuksen lähdeaineistona toimi osaotos Pohjois-Suomen asukkaista koostuvasta syntymäkohortti 1966:sta (KOHO66), jossa mukana olevien henkilöiden polvinivelet olivat röntgenkuvatut aiemmin. Röntgenkuvat luettiin ja luokiteltiin kahden tutkijan toimesta järjestelmällisesti sekä tulokset kirjattiin ylös. Molemmat tutkijat olivat lääketieteen opiskelijoita, joilla ei ollut juurikaan aikaisempaa kokemusta natiiviröntgenkuvien tulkinnasta. Röntgenkuvista oli käytettävissä yleensä sekä anteroposteriorinen että posteroanteriorinen projektio ja tulkinnassa pyrittiin käyttämään ensisijaisesti sitä projektiota, jossa nivelrako oli enemmän auki silmämääräisesti tarkasteltuna.

Ennen varsinaisen tutkimusaineiston luokittelua molemmat perehtyivät OARSI-atlakseen soveltuvilta osin sekä luokittelivat harjoitusmielessä pienemmän ns. konsensusaineiston. Konsensusaineisto oli valmiiksi luokiteltu kahden kokeneen ammattilaisen toimesta, joiden luennan synteesinä syntyneeseen tulokseen tutkijat pystyivät alussa vertaamaan omia tuloksiaan. Vaiheen tarkoituksena oli omaksua OARSI-luokitusjärjestelmä riittävällä tasolla sekä löytää merkittävät poikkeamat tuloksissa suhteessa konsensusaineistoon.

Varsinaisen tutkimusaineisto koostui noin kahdesta tuhannesta KOHO66-henkilöstä, jotka satunnaistettiin kahteen yhtä suureen noin tuhannen henkilön joukkoon tutkijoita varten. Alkuperäisessä joukossa oli kuitenkin osa samoja henkilöitä ts. redundanssia, joka mahdollisti luokittelun tutkijoiden välisen sekä tutkijan sisäisen toistettavuuden arvioinnin myöhemmässä vaiheessa. Osa tutkimusaineiston röntgenkuvista puuttui tai röntgenkuvissa oli jokin tekijä, esimerkiksi tekonivel, joka laski analysoitujen kuvien määrää. Molempien tutkijoiden luokiteltavaksi jäi lopulta noin 1900 polvea, joka käsittää siten noin 950 henkilöä. Luokittelun tuloksena saatu data analysoitiin tilastollisesti käyttäen vapaasti saatavana olevaa GNU R-ohjelmistoa sekä siihen saatavia *irr*-, *psych*- sekä *epiR*-laajennospaketteja, jotka mahdollistivat haluttujen metriikoiden laskennan. R on varsinaisesti ohjelmointikieli, jota käytetään runsaasti tilastollisessa laskennassa.

3.2. Tutkimuksessa käytetyt statistiikat

3.2.1 Kappa-statistiikka

Hyvin yleinen tapa arvioida lääketieteessä luokittelun välistä tai sisäistä toistettavuutta (tai sen luotettavuutta) on määrittää Kappa-statistiikka, jota käytettiin tässäkin tutkimuksessa (Cohen 1968, Hallgren 2012, McHugh 2012). Kappa-statistiikasta voidaan käyttää painottamatonta tai painotettua laskentatapaa. Painotettu kappa sopii paremmin ordinaaliasteikollisille muuttujille (Taulukko 1), jolloin eroavaisuuksia painotetaan halutuilla painokertoimilla. Usein tämä painotus on neliöllinen. Kappa-statistiikka ottaa huomioon sattuman vaikutuksen tulokseen (havaittu tulos vs. odotettu tulos). Tällainen vaikutus syntyy esimerkiksi silloin, kun tutkija luokittelee muuttujan puhtaasti arvaamalla. Kappa-kertoimen arvon tulkinnalle löytyy kirjallisuudesta erilaisia vaihtoehtoja (Sim ja Wright 2005, McHugh 2012, Flight ja Julious 2015) ja kapan absoluuttisen arvon tulkinta ei ole täysin yksiselitteistä. Taulukko 3 esittää kuitenkin yhtä tallaista vaihtoehtoa. Kappa voi saada arvoja väliltä −1.0–1.0, mutta yleensä sen arvo asettuu välille 0.0–1.0.

Taulukko 3. Kappa-kertoimen arvot ja niiden tulkinta (Flight ja Julious 2015).

Kappa	Toistettavuus
< 0.20	Heikko
0.21–0.40	Välttävä
0.41–0.60	Kohtalainen
0.61–0.80	Hyvä
0.81–1.00	Erittäin hyvä

Vaikka kappa-statistiikka on laajasti käytetty, voidaan senkin soveltuvuutta kliiniseen työhön kritisoida (de Vet HC ym. 2013) sen ollessa erityisen herkkä tutkittavan ilmiön prevalenssin sekä tutkijoiden henkilökohtaisen biaksen aiheuttamille vaikutuksille (Sim ja Wright 2005, Flight ja Julious 2015). Riippuukin lähtökohdista ja kysymyksen asettelusta, onko suhteelliseen luotettavuuteen liittyvä kappa-statistiikka vai esimerkiksi helpommin ymmärrettävä prosentuaalinen yksimielisyys mielekkäämpi tapa karakterisoida luokittelun tulosta ja sen luotettavuutta.

3.2.2 Luokan sisäinen korrelaatiokerroin

Luokan sisäinen korrelaatiokerroin (ICC) on usein käytetty vaihtoehto neliöllisesti painotetun kappa-statistiikan sijaan. Luokan sisäinen korrelaatio sopii tässäkin tutkimuksessa käytetyille ordinaaliasteikollisille muuttujille sen ottaessa huomioon tutkijoiden luokittelun eroavaisuuksien magnitudin sekä systemaattisen eron tutkijoiden luokittelun välillä (Hallgren 2012). Ennen ICC:n laskentaa täytyy kuitenkin valita haluttu tai sopiva variantti sen eri muodoista. Tämä riippuu siitä, onko tutkijat on valittu satunnaisesti erikseen jokaista luokiteltavaa tapausta varten, halutaanko arvon kuvastavan luokitusten konsistenttiutta vai absoluuttista yksimielisyyttä, käsitelläänkö luokitusarvoja yksittäisinä arvoina vai keskiarvoina sekä siitä, onko tutkijat valittu satunnaisesti suuresta joukosta samankaltaisia tutkijoita. ICC voi saada arvoja väliltä −1.0–1.0.

3.2.3 Prevalenssi- ja bias-korjattu kappa

Kappa-statistiikkaan liittyvän prevalenssi- ja bias-vaikutuksen takia, on kehitetty modifioituja ja vaihtoehtoisia statistiikoita. Yksi tällainen on prevalenssi- ja bias-modifioitu kappa (Byrt ym. 1993), josta käytetään yleisesti lyhennettä PABAK (prevalence-adjusted bias-adjusted kappa). PABAK sopii dikotomisille muuttujille, jollaisia tämän tutkimuksen yhteydessä lähimpänä ovat

eri OARSI-kriteerit (Taulukko 2) sekä näistä kriteereistä muodostettu lopullinen diagnoosi. Voidaan osoittaa, että kappa-statistiikka antaa pieniä lukuarvoja, vaikka havaittu yksimielisyys olisi kohtuullinen, mutta marginaaleissa on asymmetriaa tai epätasapainoa (Flight ja Julious 2015). Tällaisessa tilanteessa PABAK voi olla järkevä vaihtoehto perinteisemmälle kapalle.

3.2.4 Prevalenssi- ja bias-indeksit

Prevalenssi-indeksi (PI) ja bias-indeksi (BI) voidaan laskea helposti dikotomiselle muuttujalle kontingenssitaulujen avulla. Niiden avulla voidaan arvioida prevalenssin ja biaksen vaikutusta kappa-arvoon (Sim ja Wright 2005, Flight ja Julious 2015). Suuri PI nostaa kapan laskennassa odotetun tuloksen todennäköisyyttä ja pienentää siten kapan arvoa. Suuri BI-arvo vaikuttaa kappaa nostavasti. Ottamalla huomioon PI- ja BI-arvot, voidaan kappaa arvioida kriittisemmin.

3.2.5 Kendallin järjestyskorrelaatiokerroin

Kendallin järjestyskorrelaatiokerroin (tai Kendallin tau) on ei-parametrinen ordinaaliasteikollisille muuttujille sopiva riippuvuuden mittari (Kendall 1938). Se ei oleta muuttujien jakaumasta mitään ja se sopiikin tilanteisiin, jos muuttujat eivät välttämättä ole normaalijakautuneita. Kendallin järjestyskorrelaatiokerroin mittaa kahden muuttujan välistä assosiaatiota kuten korrelaatiokertoimet yleensäkin. Tämän tutkimuksen kontekstissa se tarkoittaa sitä, että tutkijoiden muuttujalle antamien luokitteluarvojen ei tarvitse olla täsmälleen samoja hyvän korrelaatioarvon saamiseksi. Siten arvoa ei voi käyttää absoluuttisen yksimielisyyden mittarina kuten vaikkapa prosentuaalista yksimielisyyttä. Kendallin järjestyskorrelaatiokertoimen arvo voi olla väliltä −1.0–1.0.

3.2.6 Prosentuaalinen yksimielisyys

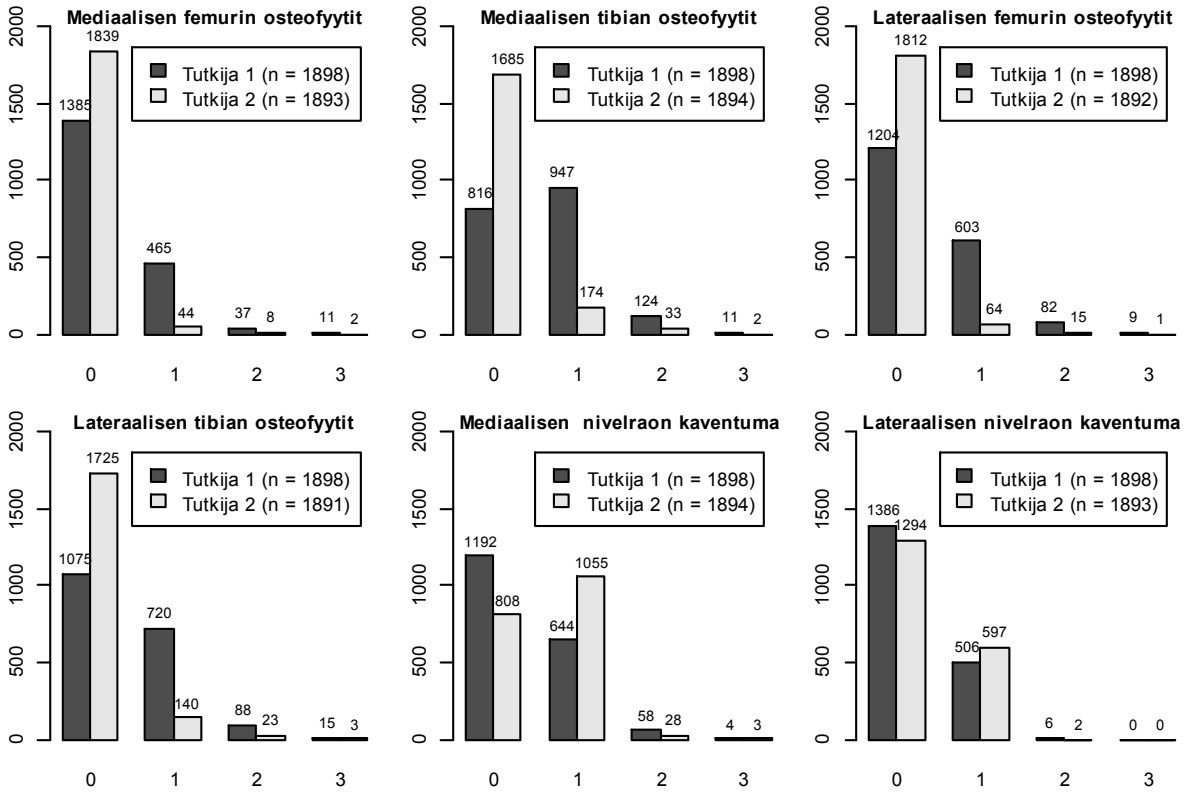
Kaikkein helpoin ja yksinkertaisin toistettavuuden mittari lienee kuitenkin havaittu yksimielisyys, joka voidaan ilmaista prosentteina. Se on yksinkertaisesti tutkijoiden yhtäpitävien luokitusten osuus kaikista luokituksista. Täytyy kuitenkin muistaa, että tällaisenkin yksinkertaisen mittarin antama tulos voi olla harhaanjohtava. Esimerkiksi dikotomisen muuttujan (ilmiö havaittavissa tai ei havaittavissa) luokittelussa saavutetaan 50 prosentin yksimielisyys, vaikka molemmat tutkijat puhtaasti arvaisivat kaikki luokittelunsa (Hallgren 2012). Prosentuaalisesta yksimielisyydestä saadaan kappa-arvon laskennassa käytettävä havaitun tuloksen todennäköisyys.

4. TULOKSET

Luokittelun tulosten toistettavuutta analysoitiin tutkijoiden välisesti että tutkijoiden sisäisesti käyttäen aikaisemmin esitettyjä tilastollisia menetelmiä ja saadut tulokset esitellään seuraavaksi.

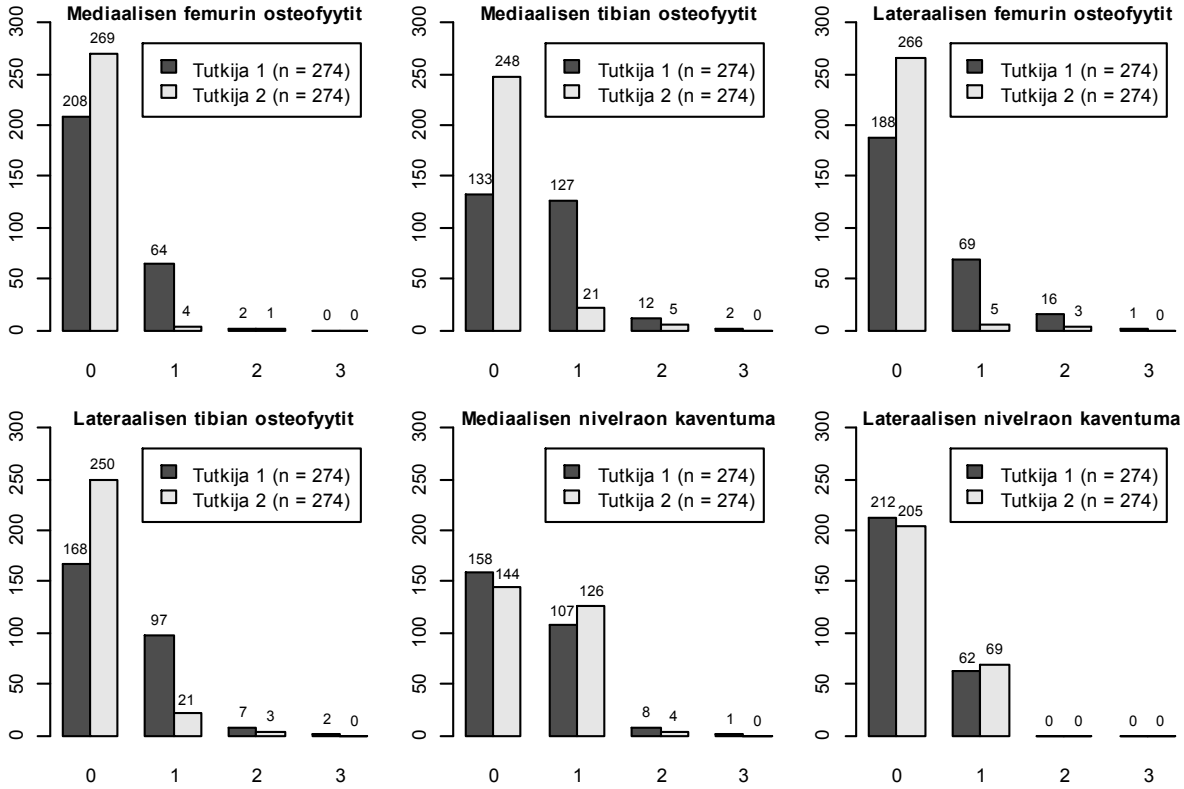
4.1. Tutkijoiden välinen toistettavuus

Kuvio 1 esittää graafisena pylväsdiagrammina, kuinka tutkijoiden antamat OARSI-luokat ovat jakautuneet koko heidän omassa osaotoksessaan eri tekijöitä luokiteltaessa. Molemmat tutkijat ovat luokitelleet varsin tarkkaan saman määrän polvia eli hiukan alle 1900 kappaletta. Kuvioista voidaan heti nähdä, että kahta korkeinta OARSI-luokkaa (2–3) esiintyy hyvin vähän suhteessa kahteen alimpaan (0–1). Tämän perusteella luokkien jakauma on varsin vino molemmilla tutkijoilla. Edelleen voidaan, todeta että tutkija 1 on käyttänyt toisiksi alinta luokkaa (1) enemmän kuin tutkija 2 osteofyyttimuodostusta luokitellessaan. Tällöin tutkijan 2 jakauman vino luonne korostuu entistä enemmän, koska luokitus painottuu nyt pääasiassa alimpaan luokkaan (0). Nivelraon kaventuman luokittelussa tutkijoiden välinen ero luokkien käytössä on selkeästi pienempi, mutta siinäkin korkeimpien luokkien (2–3) käyttö on molemmilla hyvin vähäistä verrattuna kahteen alimpaan (0–1).



Kuvio 1. OARSI-luokkien kappalemääräinen jakautuminen tutkijoiden omissa osaotoksissa luokiteltavien tekijöiden suhteen.

Kuvio 2 esittää vastaavasti OARSI-luokkien jakautumisen, mutta tällä kertaa tutkijoiden yhteisessä osaotoksessa. Tässä tapauksessa molemmat tutkijat ovat luokitelleet samat polvinivelet, joita on ollut yhteensä 274 kappaletta. Yhteisen osaotoksen jakaumaa esittävästä kuviosta voidaan nähdä, että jakaumien muodot ovat pääpiirteissään samankaltaiset kuin tutkijoiden omienkin osaotoksien jakaumissa, joten aikaisemmin tehdyt havainnot ja johtopäätökset pätevät myös tässä tapauksessa.



Kuvio 2. Eri luokkien kappalemääräinen jakautuminen tutkijoiden yhteisessä osaotoksessa luokiteltavien tekijöiden suhteen.

Taulukko 4 esittää kontingenssitauluja ristiintaulukoinnin jälkeen, jotka muodostettiin tutkijoiden yhteisen osaotoksen perusteella. Kontingenssitaulujen perusteella voidaan nähdä yksityiskohtaisemmin, kuinka tutkijat ovat käyttäneet luokitusarvoja suhteessa toisiinsa. Vasemmalta ylhäältä oikealle alas menevällä diagonaalilla näkyvät niiden tapausten lukumäärät, joissa molemmat tutkijat ovat olleet täysin samaa mieltä luokituksesta. Kunkin taulun alapuolella on näkyvissä myös Bhapkarin testin tulos, jota voidaan käyttää marginaalisen homogeenisuuden arvioinnissa eli onko rivi- ja sarakemarginaaleissa tilastollisesti merkittäviä eroja. Bhapkarin testin avulla voidaan selvittää tässä tapauksessa, onko tutkijoiden tavassa käyttää eri OARSI-luokkia merkittäviä eroavaisuuksia. Kuten kontigenssitauluja silmämääräisestikin tutkimalla havaitaan, vahvistaa myös Bhapkarin testin tulos sen, että lähes kaikissa luokitelluissa kohteissa havaitaan tilastollisesti merkittävä ero tutkijoiden välillä ($p < 0.0001$). Ainoastaan lateraalisen nivelraon kaventuman kohdalla voidaan havaita p-arvon olevan suuri ($p = 0.369$), jolloin tutkijoiden tavassa käyttää eri luokkia ei voida katsoa olevan tilastollisesti merkittävää eroa.

Tässä vaiheessa voitaneen todeta, että tutkimusaineistossa piilee potentiaalinen riski sekä prevalenssi- että bias-problematiikkaan, joihin vaikuttavat nyt käytettyjen luokkien vinot jakaumat yhdistettynä tutkijoiden erilaiseen tapaan käyttää luokkia.

Taulukko 4. Yhteisen osaotoksen luokittelusta muodostetut kontingenssitaulut.

	A. Mediaalisen femurin osteofyytit				**B.** Mediaalisen tibian osteofyytit			
	Tutkija 2				Tutkija 2			
Tutkija 1	0	1	2	3	0	1	2	3
0	206	2	0	0	133	0	0	0
1	62	1	1	0	107	20	0	0
2	1	1	0	0	8	1	3	0
3	0	0	0	0	0	0	2	0
	Bhapkar: $\chi^2 = 72.4$, $df = 2$, $p < 0.0001$				*Bhapkar:* $\chi^2 = 205$, $df = 3$, $p < 0.0001$			
	C. Lateraalisen femurin osteofyytit				**D.** Lateraalisen tibian osteofyytit			
	Tutkija 2				Tutkija 2			
Tutkija 1	0	1	2	3	0	1	2	3
0	186	2	0	0	161	6	1	0
1	67	2	0	0	85	12	0	0
2	13	1	2	0	4	3	0	0
3	0	0	1	0	0	0	2	0
	Bhapkar: $\chi^2 = 104$, $df = 3$, $p < 0.0001$				*Bhapkar:* $\chi^2 = 98.5$, $df = 3$, $p < 0.0001$			
	E. Mediaalisen nivelraon kaventuma				**F.** Lateraalisen nivelraon kaventuma			
	Tutkija 2				Tutkija 2			
Tutkija 1	0	1	2	3	0	1	2	3
0	119	39	0	0	178	34	0	0
1	25	82	0	0	27	35	0	0
2	0	5	3	0	0	0	0	0
3	0	0	1	0	0	0	0	0
	Bhapkar: $\chi^2 = 9.37$, $df = 3$, $p = 0.0247$				*Bhapkar:* $\chi^2 = 0.806$, $df = 1$, $p = 0.369$			

df: vapausaste

Taulukko 5 esittää kontingenssitaulut lasketuille OARSI-kriteereille sekä lopulliselle kriteereistä määräytyvälle nivelrikkodiagnoosille. Näkyvissä ovat vastaavasti Bhapkarin testin tulokset, jotka ovat kaikissa tapauksissa tilastollisesti merkittäviä p-arvon ollessa suurimmillaankin 0.024. Tästä voidaan edelleen päätellä, että tutkijoiden välillä on eroavaisuuksia kriteeri- ja diagnoositasolla. Huomio voidaan erityisesti kiinnittää itse diagnoosiin, jossa tutkijan 1 positiivisten diagnoosien summa on 136 ja tutkijan 2 positiivisten diagnoosien summa ainoastaan 16. Nivelrikon prevalenssissa on siten merkittävä ero tutkijoiden välillä. Erityisesti tutkijan 1 tuloksen ollessa nyt jopa noin 50 % (136/274), on se varsin poikkeava ja odottamaton tutkittua KOHO66-aineistoa ajatellen. Tällainen tulos herättää terveen epäilyn tutkijan 1 luokittelun onnistumisesta. Tarkasteltaessa tutkijan 1 yksittäisiä kriteerejä, voidaan huomata, että merkittävä osa positiivisista diagnooseista on tultava OARSI-kriteereistä 2 ja 3. Kriteeri 1 (nivelraon kaventuma ≥ 2) tuottaa positiivisen diagnoosin ainoastaan viidessä tapauksessa. Jo aikaisemmin

huomattiin, että tutkija 1 on käyttänyt osteofyyttiluokituksessa huomattavasti enemmän luokkaa 1 verrattuna toiseen tutkijaan, joka todennäköisesti johtaa nyt kriteerien 2 ja 3 kautta nivelrikon ylisuureen prevalenssiin tutkimusaineistossa.

Taulukko 5. Yhteisen osaotoksen luokittelun perusteella laskettujen OARSI-kriteereiden ja diagnoosin kontingenssitaulut.

	A. OARSI-kriteeri 1 Tutkija 2		**B.** OARSI-kriteeri 2 Tutkija 2		**C.** OARSI-kriteeri 3 Tutkija 2		**D.** DG Tutkija 2	
Tutkija 1	0	1	0	1	0	1	0	1
0	265	0	177	0	196	4	138	0
1	5	4	89	8	68	6	120	16
	A. *Bhapkar:* $\chi^2 = 5.09$, $df = 1$, $p = 0.0240$				***B.*** *Bhapkar:* $\chi^2 = 132$, $df = 1$, $p < 0.0001$			
	C. *Bhapkar:* $\chi^2 = 71.8$, $df = 1$, $p < 0.0001$				***D.*** *Bhapkar:* $\chi^2 = 214$, $df = 1$, $p < 0.0001$			

OARSI-kriteeri 1: Nivelraon kaventuman luokka ≥ 2
OARSI-kriteeri 2: Osteofyyttiluokkien summa ≥ 2
OARSI-kriteeri 3: Luokan 1 nivelraon kaventuma yhdessä luokan 1 osteofyytin kanssa
DG: Diagnoosi, positiivinen mikäli OARSI-kriteeri 1, 2 tai 3 täyttyy
df: vapausaste

Taulukko 6 esittää tutkijoiden yhteisestä osaotoksesta määritettyjä luokittelijoiden väliseen toistettavuuteen liittyviä tunnuslukuja 95 % luottamusväleineen. Tunnusluvut laskettiin osteofyyttiluokituksille, nivelrakoluokituksille, OARSI-kriteereille sekä lopulliselle diagnoosille. Osteofyytti- sekä nivelrakoluokitukset olivat neliportaisella ordinaaliasteikolla loppujen ollessa dikotomisia. Kaikille laskettiin samat tunnusluvut, vaikka luokan sisäisen korrelaation laskentaa käytetäänkin yleensä useampiportaiselle ordinaali- tai välimatka-asteikollisille muuttujille. Dikotomisen asteikon omaaville tekijöille laskettiin lisäksi PABAK-, PI- sekä BI-arvot. Neliportaisten ordinaaliasteikollisten muuttujien kappa-arvo on lisäksi neliöllisesti painotettu erotuksena dikotomisten muuttujien painottamattomaan kappa-arvoon. Luokan sisäisestä korrelaatiosta määritettiin ICC(3,1) -arvo. Tämä merkintätapa tarkoittaa tilannetta, jossa kiinteä joukko arvioijia (ei yleistystä suurempaan samanlaiseen joukkoon tutkijoita) luokittelee jokaisen tutkimuskohteen ja luotettavuus lasketaan yksittäisistä mittauksista (eikä niiden keskiarvoista).

Tarkasteltaessa osteofyyttiluokitusten tuloksia, voidaan havaita prosentuaalisen yksimielisyyden vaihtelevan eri tapauksissa välillä 56.9–75.5 % sekä kappa-arvon välillä 0.091–0.28. Prosentuaalinen yksimielisyys on huonoimmillaan varsin heikko ollen kuitenkin parhaimmillaan kohtuullinen. Kappa-arvo saavuttaa parhaimmillaankin vain välttävän tason (Taulukko 3). Heikoin kappa-arvo 0.091 saavutetaan kuitenkin silloin kuin prosentuaalinen yksimielisyys on korkeimmillaan 75.5 % mediaalisen femurin tapauksessa sekä paras kappa-arvo 0.28 saavutetaan, kun prosentuaalinen yksimielisyys on pienimmillään 56.9 % mediaalisen tibian tapauksessa.

Muut korrelaatiometriikat (Kendall ja ICC(3,1)) eivät poikkea merkittävissä määrin suhteessa kappaan.

Taulukko 6. Yhteisen osaotoksen nivelrikkomuuttujien luokittelun sekä niistä laskettujen OARSI-kriteerien ja diagnoosin tutkijoiden välisen toistettavuuden tilastolliset tunnusluvut.

	A. Osteofyytit			
	Mediaalinen		Lateraalinen	
	Femur	Tibia	Femur	Tibia
Kappa	0.091 (−0.021–0.20)	0.28 (0.16–0.41)	0.18 (0.028–0.34)	0.24 (0.087–0.40)
Kendall	0.13 (0.0096–0.24)	0.35 (0.24–0.45)	0.20 (0.084–0.31)	0.24 (0.12–0.35)
ICC(3,1)	0.011 (−0.069–0.23)	0.40 (0.30–0.50)	0.23 (0.12–0.34)	0.30 (0.19–0.40)
Yksimielisyys [%]	75.5	56.9	69.3	63.1
	B. Nivelraon kaventuma			
	Mediaalinen		Lateraalinen	
Kappa	0.58 (0.48–0.68)		0.39 (0.26–0.51)	
Kendall	0.55 (0.46–0.63)		0.39 (0.28–0.49)	
ICC(3,1)	0.58 (0.50–0.65)		0.39 (0.28–0.49)	
Yksimielisyys [%]	74.5		77.7	
	C. OARSI-kriteeri			**D.** DG
	1	2	3	
Kappa	0.61 (0.29–0.92)	0.10 (0.036–0.17)	0.084 (−0.0041–0.17)	0.12 (0.063–0.17)
Kendall	0.66 (0.59–0.72)	0.23 (0.12–0.34)	0.14 (0.027–0.26)	0.25 (0.14–0.36)
ICC(3,1)	0.61 (0.53–0.68)	0.15 (0.029–0.26)	0.10 (−0.015–0.22)	0.19 (0.076–0.30)
PABAK	0.96 (0.92–0.99)	0.35 (0.23–0.46)	0.47 (0.36–0.58)	0.12 (0.0021–0.24)
PI	0.95 (0.93–0.98)	0.62 (0.56–0.68)	0.69 (0.64–0.75)	0.45 (0.38–0.51)
BI	−0.018 (−0.044–0.0072)	−0.32 (−0.38–−0.26)	−0.23 (−0.29–−0.18)	−0.44 (−0.50–−0.37)
Yksimielisyys [%]	98.2	67.5	73.7	56.2

Kappa: Neliöllisesti painotettu arvo paitsi kohdissa C ja D painottamaton, (95 % luottamusväli)
Kendall: Kendallin järjestyskorrelaatiokerroin, (95 % luottamusväli)
ICC(3,1): Luokan sisäinen korrelaatiokerroin, (95 % luottamusväli)
PABAK: Prevalenssi- ja bias-korjattu kappa, (95 % luottamusväli)
PI: Prevalenssi-indeksi, (95 % luottamusväli)
BI: Bias-indeksi, (95 % luottamusväli)
OARSI-kriteeri 1: Nivelraon kaventuman luokka ≥ 2
OARSI-kriteeri 2: Osteofyyttiluokkien summa ≥ 2
OARSI-kriteeri 3: Luokan 1 nivelraon kaventuma yhdessä luokan 1 osteofyytin kanssa
DG: Diagnoosi, positiivinen mikäli OARSI-kriteeri 1 tai 2 tai 3 täyttyy
n = 274

Nivelraon luokituksen tunnuslukuja tarkasteltaessa voidaan niiden havaita olevan hiukan parempia kuin osteofyyttiluokituksessa arvojen ollessa nyt 0.58 ja 0.39. Mediaalisen kaventuman tapauksessa kappa saavuttaa siis jo kohtalaisen tason. Prosentuaalinen yksimielisyys on kohtalaista tasoa arvojen ollessa nyt 74.5 % ja 77.7 %. Korrelaatiometriikat tuottavat identtisen tuloksen suhteessa kappaan.

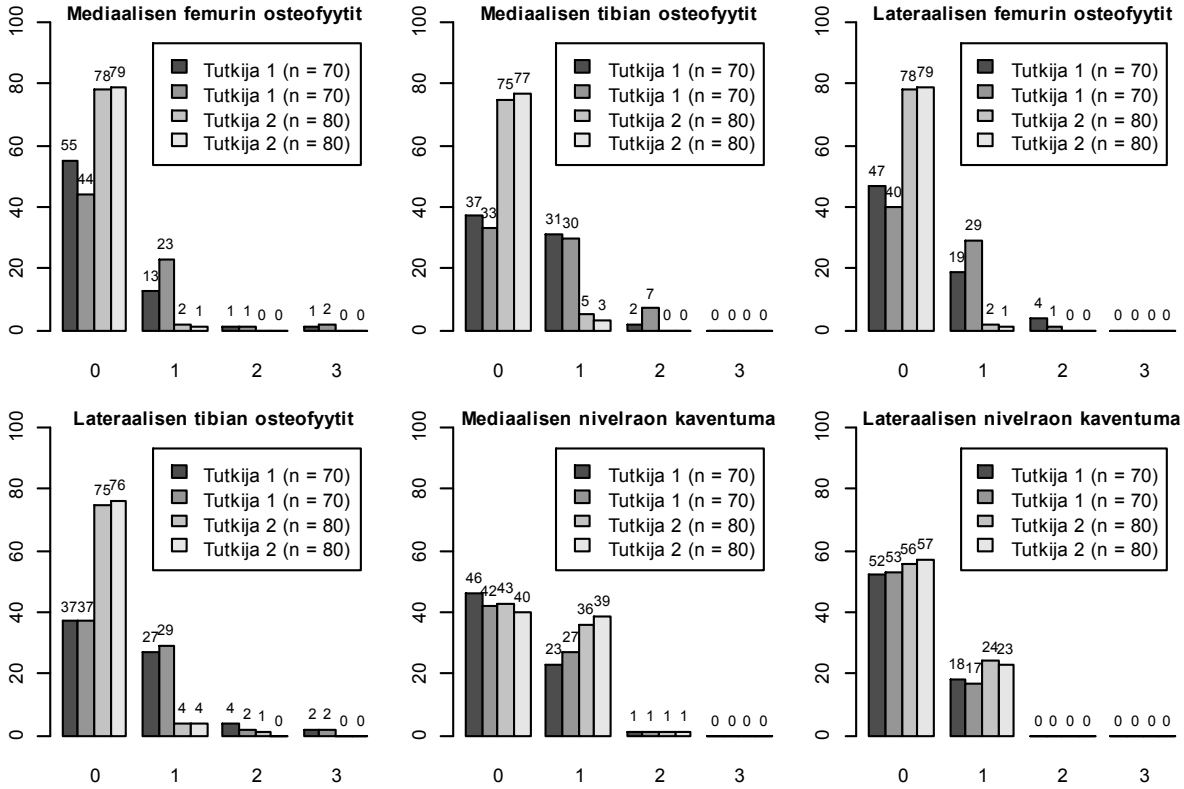
OARSI-kriteerien tunnuslukuja tarkasteltaessa huomataan kriteerin 1 (puhtaasti nivelrakokriteeri) prosentuaalisen yksimielisyyden olevan erittäin hyvä 98.2 %, mutta kappa saavuttaa vain kohtalaisen arvon 0.61. ICC(3,1) ja Kendallin järjestyskorrelaatiokerroin ovat linjassa kapan

kanssa. Sen sijaan prevalenssi- ja bias-korjattu PABAK saa nyt korkean arvon 0.96. Kriteerien 2 ja 3 tapauksessa kapan sekä korrelaatiometriikoiden arvot jäävät heikolle tasolle. Ainoastaan PABAK saavuttaa näissä jonkin verran korkeammat arvot. Prevalenssi- indeksin arvo on korkea kaikkien OARSI-kriteerien tapauksessa ja todettavissa kohonneeksi myös diagnoosin tasolla. Bias-indeksin arvo on kriteerin 1 tapauksessa hyvin pieni, mutta kriteerien 2 ja 3 sekä diagnoosin kohdalla kohonnut. Prevalenssi-indeksillä on nyt kappaa laskeva vaikutus kriteerien 1–3 tapauksessa ja saatu PABAK-arvo onkin näissä tapauksissa korkeampi. Sen sijaan lopullisen diagnoosin tapauksessa PABAK-arvokin jää varsin heikoksi.

Kaiken kaikkiaan kohtalainen kappa (≥ 0.41) saavutetaan nyt ainoastaan kahdessa kohdassa kymmenestä. Lopullista diagnoosia tarkasteltaessa, voidaan huomata kaikkien toistettavuutta mittaavien metriikoiden jäävän valitettavan mataliksi. Helpoimmin ymmärrettävä prosentuaalinen yksimielisyys on nyt ainoastaan 56.2 %, jota ei voitane pitää hyvänä saavutuksena eikä liioin kappa- tai PABAK-arvoa niiden molempien ollessa 0.12 ja kuvastaessa siten heikkoa toistettavuutta. Muut metriikat eivät tuota merkittävästi näistä poikkeavia tuloksia. Mielenkiintoista on huomata, että vaikka PABAK saakin parempia arvoja kuin kappa OARSI-kriteerien kohdalla, ei samanlaista ilmiötä ole enää havaittavissa kuitenkaan diagnoosin tasolla. Tämän tutkimuksen perusteella OARSI-nivelrikkoluokituksen tutkijoiden välisen toistettavuuden ei voi sanoa olevan hyvä (diagnoosin tasolla), vaikka yksittäisten luokiteltavien muuttujien kohdalla päästiinkin välillä kohtalaisiin tuloksiin.

4.2. Tutkijoiden sisäinen toistettavuus

Kuvio 3 esittää, kuinka tutkijoiden luokitukset ovat jakautuneet heidän sisäisen toistettavuuden laskentaan käytetyissä osaotoksissa. Tutkijan 1 aineistossa toistuvia polvia on ollut 70 kappaletta ja tutkijalla 2 vastaavasti 80 kappaletta. Kuviosta voidaan nähdä, että tutkijalla 2 on ollut vähemmän kappalemääräistä vaihtelua verrattuna tutkijaan 1 eri luentakertojen kesken. Tutkijalla 2 vaihtelu on ollut suurimmillaankin vain 3 kappaletta (mediaalisen nivelraon kaventuma) huolimatta jonkin verran korkeammasta luokiteltujen polvien määrästä. Tutkijalla 1 vaihtelu on korkeimmillaan 11 kappaletta (lateraalisen femurin osteofyytit). Edelleen voidaan havaita, että tässäkin tapauksessa tutkija 1 on käyttänyt omalla kohdallaan enemmän luokkaa 1 osteofyyttien luokittelussa kuin tutkija 2.



Kuvio 3. Eri luokkien kappalemääräinen jakautuminen tutkijoiden sisäisen toistettavuuden osaotoksissa luokiteltavien tekijöiden suhteen.

Taulukko 7 esittää tutkijoiden omista osaotoksista määritettyjä luokittelun sisäisen toistettavuuden tunnuslukuja 95 % luottamusväleineen. Sisäisen toistettavuuden tapauksessa laskettiin kappa, Kendallin järjestyskorrelaatiokerroin, ICC(3,1) sekä prosentuaalinen yksimielisyys.

Taulukko 7. Tutkijoiden omien osaotoksien nivelrikkomuuttujien luokittelun sekä niistä laskettujen OARSI-kriteerien ja diagnoosin sisäinen toistettavuuden tilastolliset tunnusluvut.

	A. Osteofyytit			
	Mediaalinen femur		Mediaalinen tibia	
	Tutkija 1	Tutkija 2	Tutkija 1	Tutkija 2
Kappa	0.69 (0.49–0.90)	0.66 (0.04–1.3)	0.60 (0.44–0.77)	0.48 (0.040–0.91)
Kendall	0.57 (0.39–0.71)	0.70 (0.57–0.80)	0.63 (0.46–0.75)	0.49 (0.31–0.64)
ICC(3,1)	0.72 (0.59–0.82)	0.66 (0.52–0.77)	0.62 (0.45–0.74)	0.48 (0.29–0.63)
Yksimielisyys [%]	75.7	98.8	74.3	95.0
	B. Osteofyytit			
	Lateraalinen femur		Lateraalinen tibia	
	Tutkija 1	Tutkija 2	Tutkija 1	Tutkija 2
Kappa	0.27 (0.051–0.50)	0.66 (0.040–1.3)	0.80 (0.67–0.94)	0.65 (0.37–0.93)
Kendall	0.30 (0.071–0.50)	0.70 (0.57–0.80)	0.70 (0.55–0.80)	0.65 (0.51–0.76)
ICC(3,1)	0.28 (0.045–0.48)	0.66 (0.52–0.77)	0.80 (0.70–0.87)	0.65 (0.51–0.76)
Yksimielisyys [%]	62.9	98.8	80.0	95.0
	C. Nivelraon kaventuma, mediaalinen		**D.** Nivelraon kaventuma, lateraalinen	
	Tutkija 1	Tutkija 2	Tutkija 1	Tutkija 2
Kappa	0.62 (0.42–0.83)	0.70 (0.55–0.86)	0.58 (0.36–0.80)	0.61 (0.42–0.80)
Kendall	0.59 (0.42–0.73)	0.69 (0.55–0.79)	0.58 (0.40–0.72)	0.61 (0.45–0.73)
ICC(3,1)	0.63 (0.46–0.75)	0.71 (0.58–0.80)	0.58 (0.40–0.72)	0.61 (0.45–0.73)
Yksimielisyys [%]	80.0	83.8	84.3	83.8
	E. OARSI-kriteeri 1		**F.** OARSI-kriteeri 2	
	Tutkija 1	Tutkija 2	Tutkija 1	Tutkija 2
Kappa	1.0 (–)	1.0 (–)	0.25 (0.035–0.46)	0.00 (–)
Kendall	1.0 (–)	1.0 (–)	0.27 (0.033–0.47)	– (–)
ICC(3,1)	1.0 (–)	1.0 (–)	0.26 (0.033–0.47)	0.00 (−0.22–0.22)
PABAK	1.0 (0.85–1.0)	1.0 (0.87–1.0)	0.26 (0.0095–0.48)	0.95 (0.83–0.99)
PI	0.97 (0.93–1.0)	0.98 (0.94–1.0)	0.20 (0.050–0.35)	0.98 (0.94–1.0)
BI	0.0 (−0.039–0.039)	0.0 (−0.034–0.034)	0.17 (0.012–0.33)	−0.025 (−0.059–0.0092)
Yksimielisyys [%]	100	100	62.9	97.5
	G. OARSI-kriteeri 3		**H.** DG	
	Tutkija 1	Tutkija 2	Tutkija 1	Tutkija 2
Kappa	0.42 (0.19–0.65)	0.79 (0.40–1.2)	0.27 (0.058–0.48)	0.85 (0.56–1.1)
Kendall	0.42 (0.21–0.60)	0.81 (0.72–0.87)	0.29 (0.055–0.49)	0.86 (0.79–0.91)
ICC(3,1)	0.42 (0.21–0.60)	0.80 (0.70–0.86)	0.29 (0.056–0.49)	0.85 (0.78–0.90)
PABAK	0.51 (0.28–0.70)	0.98 (0.86–1.0)	0.26 (0.0095–0.48)	0.98 (0.86–1.0)
PI	0.41 (0.27–0.56)	0.94 (0.88–0.99)	−0.11 (−0.27–0.038)	0.91 (0.85–0.98)
BI	0.071 (−0.079–0.22)	−0.013 (−0.066–0.041)	0.17 (0.0093–0.33)	−0.013 (−0.076–0.050)
Yksimielisyys [%]	75.7	98.8	62.9	98.8

Kappa: Neliöllisesti painotettu arvo paitsi kohdissa E–H painottamaton, (95 % luottamusväli)
Kendall: Kendallin järjestyskorrelaatiokerroin, (95 % luottamusväli)
ICC(3,1):Luokan sisäinen korrelaatiokerroin, (95 % luottamusväli)
PABAK: Prevalenssi- ja bias-korjattu kappa, (95 % luottamusväli)
PI: Prevalenssi-indeksi, (95 % luottamusväli)
BI: Bias-indeksi, (95 % luottamusväli)
OARSI-kriteeri 1: Nivelraon kaventuman luokka ≥ 2
OARSI-kriteeri 2: Osteofyyttiluokkien summa ≥ 2
OARSI-kriteeri 3: Luokan 1 nivelraon kaventuma yhdessä luokan 1 osteofyytin kanssa
DG: Diagnoosi, positiivinen mikäli OARSI-kriteeri 1, 2 tai 3 täyttyy
Tutkija 1: n = 70; Tutkija 2: n = 80

Vertailtaessa sisäisen toistettavuuden lukuja tutkijoiden välillä, voidaan huomata, että tutkijalla 2 prosentuaalinen yksimielisyys on alimmillaankin 83.8 % ja usein yli 95 % eli hyvin korkea. Tutkijalla 1 vaihtelua on huomattavasti enemmän arvon vaihdellessa välillä 62.9–100 %. Tämä ero tutkijoiden välillä ei kuitenkaan näy aina kappa-arvossa (tai muissakaan korrelaatiometriikoissa), vaan vaikutus on jopa päinvastainen kapan ollessa hiukan korkeampi tutkijalla 1. OARSI-kriteerin 2 tapauksessa tutkijan 2 kappa saavuttaa poikkeuksellisesti arvon nolla.

Kappa-arvon valossa tutkija 1 on saavuttanut vähintään kohtalaisen toistettavuuden tason (kappa $\geq$ 0.41) seitsemässä kohdassa kymmenessä ja hyvän tason (kappa $\geq$ 0.61) neljässä kohdassa kymmenessä. Vastaavasti tutkijalla 2 vähintäänkin kohtalainen toistettavuus on saavutettu yhdeksässä kohdassa kymmenestä ja hyvä taso kahdeksassa kohdassa kymmenestä. Näyttää siis siltä, että tutkijalla 2 sisäinen toistettavuus olisi parempi. Erityisesti lopullisen diagnoosin tasolla tutkijoiden kapan arvot poikkeavat merkittävästi toisistaan tutkijalla 1 arvon ollessa vain välttävä 0.27 ja tutkijalla 2 sen ollessa nyt erittäin hyvä 0.85. Täten OARSI-nivelrikkoluokituksen sisäinen toistettavuus näyttäisi riippuvan merkittävästi sitä tekevästä tutkijasta.

5. POHDINTA

Tämän tutkimuksena tarkoituksena oli selvittää polven OARSI-nivelrikkoluokituksen tutkijoiden välistä että tutkijoiden sisäistä toistettavuutta. Toistettavuudelle saavutettiin varsin vaihtelevia tuloksia riippuen siitä määritettiinkö luotettavuus tutkijoiden välisesti, tutkijoiden sisäisesti ja oliko tutkittavana muuttujana diagnoosiin vaikuttava osatekijä vai itse lopullinen diagnoosi.

Tutkimuksen lähdemateriaalina toimi osaotos syntymäkohortti 1966:sta. Kun tarkastellaan, kuinka tutkijat ovat käyttäneet eri luokitusarvoja, huomattiin suurempien luokitusarvojen (2–3) merkittävä vähäisyys suhteessa kahteen alimpaan (0–1). Täten tutkimusaineistoa voidaan pitää yksipuoleisena ajatellen OARSI-luokitusta ja monipuolisempi materiaali olisi ollut paikallaan. Tässä tutkimuksessa ei kuitenkaan suoraan verrattu OARSI-luokituksen paremmuutta tai huonommuutta suhteessa johonkin toiseen nivelrikon luokitusjärjestelmään. Monipuolisempi aineisto olisi ehkä ollut myös parempi tutkijoiden kannalta, jolloin eri luokkien merkitys ja käyttö olisi voinut selkiytyä heille paremmin. Tämä siitäkin huolimatta, että tutkijat luokittelivat aluksi monipuolisemman, mutta merkittävästi lukumäärältään pienemmän konsensusaineiston.

Kuva 1. Esimerkkejä polvista molempine projektioineen, joissa tutkijoiden välille on muodostunut eroja useiden luokiteltavien tekijöiden suhteen (tutkija 1–tutkija 2). Ylin kuvapari: Osteofyytit, mediaalinen femur 1–0, mediaalinen tibia 1–0, lateraalinen femur 2–0, lateraalinen tibia 1–0; nivelraon kaventuma, mediaalinen 2–1, lateraalinen 1–1. Keskimmäinen kuvapari: Osteofyytit, mediaalinen femur 1–0, mediaalinen tibia 1–0, lateraalinen femur 1–0, lateraalinen tibia 2–0; nivelraon kaventuma, mediaalinen 1–1, lateraalinen 1–1. Alin kuvapari: Osteofyytit, mediaalinen femur 1–0, mediaalinen tibia 1–0, lateraalinen femur 1–0, lateraalinen tibia 1–0; nivelraon kaventuma, mediaalinen 0–1, lateraalinen 0–0.

Edelleen tutkijoiden tavassa käyttää eri luokkia havaittiin eroja, joka huomattiin erityisesti osteofyyttiluokitusten tapauksessa. Nivelrakoluokkien kohdalla tutkijoiden luokkien käyttö suhteessa toisiinsa oli tasaisempaa. Koska käytetyt luokat painottuivat aiemmin kerrotulla tavalla kahteen alimpaan, havaittiin tutkijoiden välillä merkittävää eroa siten luokkien 0 ja 1 käytön kesken. Tutkija 1 on käyttänyt huomattavasti enemmän luokkaa 1 kuin tutkija 2 (Kuva 1). Tällä on varmastikin merkitystä, kuinka eri OARSI-kriteerit muodostuvat ja edelleen sanelevat lopulta koko diagnoosin. Näyttäisi siltä, että tutkija 1 on ollut huomattavan paljon epävarmempi kahden alimman luokan käytössä, kun taas tutkija 2 on pystynyt pitäytymään systemaattisemmin omalla luokittelulinjallaan. Tämä näkyi tutkijoiden sisäisen toistettavuuden arvioinnin kohdalla, jossa tutkija 2 saa hyvinkin korkeita prosentuaalisen yksimielisyyden arvoja sekä hyvän kappa-arvon lopulliselle diagnoosille. Tutkijalla 1 havaittiin hyvin korkea positiivisten diagnoosien osuus yhteisessä aineistossa, joka johtunee nyt suuremmasta luokan 1 käytöstä osteofyyttien tapauksessa. Tämä heijastunee nyt OARSI-kriteerien 2 ja 3 kautta itse diagnoosiin. Tässä suhteessa alkuperäinen OARSI-diagnoosin määritelmä tuntuu varsin herkältä mittarilta ajatellen alimpia luokka ja korkeampien luokkien merkitys voi tuntua epäselvältä. Toisaalta tutkijan 1 voidaan ajatella luokitelleen osteofyyttejä aivan liian helposti luokkaan 1.

Tutkijoiden kokemattomuus radiologisten nivelrikkomuuttujien tunnistamisessa ja luokittelussa saattaa selittää myös sen, miksi tutkijoiden välinen toistettavuus jäi varsin heikoksi. Tämä siitäkin huolimatta, että tutkijat kävivät ennen varsinaista luokittelutyön aloittamista läpi lyhyen perehdytysjakson. Toisaalta taas luokiteltujen polvien korkea lukumäärä on hyvä asia tutkimuksen kannalta. Edelleen suurempi tutkijoiden määrä antaisi uutta näkökulmaa ja toisi lisäluotettavuutta ajatellen tutkijoiden välistä toistettavuutta. Mielenkiintoista olisikin nähdä tämän tutkimuksen tutkijoiden luokittelu suhteessa esimerkiksi kokeneen radiologin tekemään luokitteluun.

Aikaisemmissa tutkimuksissa on jo todettu, että OARSI-järjestelmällä nivelrikkkodiagnoosien määrä on lähes kaksinkertainen verrattuna Kellgren & Lawrence -järjestelmään samassa aineistossa, suurimpien tutkijoiden välisten erojen liittyessä osteofyytteihin sekä tutkijoiden välisen toistettavuuden olevan parhaimmillaankin vain kohtalainen (Culvenor ym. 2014). Siten tällaisten tutkimusten tulokset ovat erityisen riippuvaisia käytetystä luokitusjärjestelmästä, käytetystä aineistosta sekä tutkijoista. Tämänkin tutkimuksen tulokset vahvistavat näitä havaintoja. Siten eri tutkimuksissa ilmoitettuihin polven nivelrikon prevalenssilukuhin sekä eri luokitusjärjestelmien kliinisten päätösrajojen ekvivalenssiin täytyy suhtautua varauksella.

Yleisesti käytetyn toistettavuuden mittarin eli kappa-statistiikan luonne ja käyttäytyminen tulisi ottaa huomioon tulosten tulkinnassa. Tämä tosiasia onkin kirjallisuudessa dokumentoitu ja esimerkiksi prevalenssi- sekä bias-indeksien määritys auttaa tässä suhteessa (Byrt ym. 1993, Sim ja Wright 2005, Flight ja Julious 2015). Prevalenssi- ja bias-korjattu kappa eli PABAK antoikin joissakin tapauksissa merkittävästi perinteisestä kapasta erovia ja sitä suurempia lukuarvoja tässäkin tutkimuksessa.

6. LÄHTEET

Altman RD, Gold GE. Atlas of individual radiographic features in osteoarthritis, revised. Osteoarthritis Cartilage 2007; 15: A1-56.

Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol 1993; 46: 423-9.

Cohen J. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychol Bull 1968; 70: 213-20.

Culvenor A, Engen C, Øiestad B, Engebretsen L, Risberg M. Defining the presence of radiographic knee osteoarthritis: a comparison between the Kellgren and Lawrence system and OARSI atlas criteria. Knee Surg Sports Traumatol Arthrosc 2014: 1-8.

de Vet HC, Mokkink LB, Terwee CB, Hoekstra OS, Knol DL. Clinicians are right not to like Cohen's kappa. BMJ 2013; 346: f2125.

Flight L, Julious SA. The disagreeable behaviour of the kappa statistic. Pharmaceut Statist 2015; 14: 74-8.

Hallgren KA. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. Tutor Quant Methods Psychol 2012; 8: 23-34.

Heliövaara M, Slätis P, Paavolainen P. Nivelrikon esiintyvyys ja kustannukset. Duodecim 2008; 124: 1869-74.

Jurvelin JS, Nieminen MT, Töyräs J, ym. Fysikaaliset ja kemialliset menetelmät nivelrikon varhaisessa osoittamisessa. Duodecim 2008; 124: 1885-96.

Kellgren Jh, Lawrence Js. Radiological assessment of osteo-arthrosis. Ann Rheum Dis 1957; 16: 494-502.

Kendall MG. A New Measure of Rank Correlation. Biometrika 1938; 30: 81-93.

Konttinen YT, Lindroos L, Ruuttila P, ym. Nivelrikon kliininen kuva ja hoito. Duodecim 2003; 119: 1537-44.

McHugh ML. Interrater reliability: the kappa statistic. Biochem Med 2012; 22: 276-82.

Ojala R, Arokoski J. Lonkka- ja polvinivelrikon vaikeusasteen radiologiset luokittelumenetelmät ja menetelmien toistettavuus. Duodecim 2012 (Luettu 22.10.2015). Saatavissa: http://www.kaypahoito.fi/web/kh/suositukset/suositus?id=nak05664.

Remes V, Virolainen P, Kettunen J, Miettinen H. Polven nivelrikon kirurginen hoito. Duodecim 2008; 124: 261-70.

Sim J, Wright CC. The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. Physical Therapy 2005; 85: 257-68.