TURUN YLIOPISTO
Kauppakorkeakoulu

Tiedonhallintamallin vaikutus dataprojekteihin terveysteknologia-alan pk-yrityksissä

Tietojärjestelmätieteen
pro gradu -tutkielma

Laatija:
Jarkko Rantanen

Ohjaaja(t):
FL Antti Tuomisto
KTM Tanja Vähämäki

11.5.2023
Turku

Pro gradu -tutkielma

Oppiaine: Tietojärjestelmätiede
Tekijä(t): Jarkko Rantanen
Otsikko: Tiedonhallintamallien vaikutus dataprojekteihin terveysteknologia-alan pk-yrityksissä
Ohjaaja(t): FL Antti Tuomisto, KTM Tanja Vähämäki
Sivumäärä: 69 sivua + liitteet 6 sivua
Päivämäärä: 11.5.2023

Tiivistelmä

Henkilökohtaisten terveystietojen digitalisointi on lisääntynyt huomattavasti ja niiden hyödyntäminen data-analytiikassa on yleistynyt merkittävästi. Terveystietoja kertyy massiivisiin tietojoukkoihin, joita voidaan kutsua myös Big Dataksi. Tiedonhallintamallin muodostamisella pyritään hallitsemaan massiivisia tietojoukkoja ja sen tarkoituksena on kertoa mitä päätöksiä tiedonhallinnan ja käytön varmistamiseksi tehdään ja miten lisätään datasta hyödynnettävää arvoa sekä minimoidaan siihen liittyviä kustannuksia ja riskejä. Terveystietojen keräämisessä ja käsittelyssä on runsaasti potentiaalia, mutta tiedonhallintamallit eivät ole kyenneet pysymään tämän kehityksen tahdissa.

Tutkimus keskittyi selvittämään tiedonhallintamallin roolia, pk-yrityksissä, jotka kehittävät data-analytiikkaratkaisuja terveydenhuollon ympäristöön. Tiedonhallintamallin käyttöönotto on tärkeää pk-yrityksille, jotta voidaan hallita paremmin suuria tietojoukkoja regulaatioiden ja vaatimusten mukaan sekä kehittää vaatimustenmukaisia data-analytiikkaratkaisuja. Vaatimustenmukaisuus on erityisen tärkeää terveydenhuoltoalan yhteydessä, koska pk-yrityksien tietojoukot sisältävät ihmisten terveyteen liittyviä tietoja. Tutkimuksen teoreettinen viitekehys muodostettiin kirjallisuuskatsauksen avulla ja se perustuu jo kehitettyihin tiedonhallintamalleihin, jotka ovat sopivia juuri dataprojektien ja terveystietojen hallinnan tarpeisiin. Lisäksi teoreettisesta viitekehyksestä pyrittiin tekemään sellainen, että sen osat ovat yhteensopivia CRISP-DM prosessimallin kanssa. CRISP-DM (englanniksi *CRoss Industry Standard Process for Data Mining)* on yleinen toimialasta riippumaton malli dataprojekteille, jonka avulla luodaan dataa hyödyntäviä analytiikka- ja tekoälyratkaisuja.

Tutkimuksen empiiristä osiota lähestyttiin laadullisesti ja tutkimusaineisto koostui viidestä teemahaastattelusta. Empiirisen tutkimuksen tulosten perusteella tarkennettiin vielä kirjallisuuskatsauksessa määriteltyä tiedonhallintamallin viitekehystä. Tuloksista kävi ilmi, että tiedonhallintamalli tulisi ottaa käyttöön terveysteknologia-alan pk-yrityksen ydintoiminnassa ja sen merkitys erityisesti kasvavalle yritykselle, jolla Big data -tietojoukkoja alkaa olla useita, on suuri. Tiedonhallintamallin puute muodostaa pullonkaulan kehitystyölle, jonka vuoksi dataprojektien eteneminen on hidasta. Jotta Big datan hyödyt saadaan kokonaisvaltaisemmin käyttöön, terveysteknologia-alan pk-yritykset tarvitsevat yhä parempia tiedonhallintamallin prosesseja.

Merkittävä huomio tutkimuksessa oli se, että terveysteknologia-alan asiakkaiden vaatimusten täyttäminen on jo valmiiksi vaikeaa, mutta tutkimuksessa määritelty tiedonhallintamalli auttaa jäsentämään kokonaisuuden ja osoittamaan juuri niihin kohtiin, joihin resurssit kannattaa kohdistaa ennalta. Esimerkiksi tietojen varastointiin tarkoitetun palvelimen hankinta EU:n alueelta helpottaa merkittävästi sidosryhmien ja GDPR:n vaatimusten noudattamista, mikä nopeuttaa vaatimustenmukaisuuteen kuluvaa raskasta työtä ja vapauttaa aikaa pk-yritykselle mielekkäämpiin dataprojektivaiheisiin, kuten datan mallinnusvaiheeseen.

Avainsanat: Tiedonhallintamalli, Big Data, GDPR, CRISP-DM

SISÄLLYS

KUVIOT

TAULUKOT

1 Johdanto

1.1 Motivointi

Viime vuosina henkilökohtaisten terveystietojen digitalisointi on lisääntynyt huomattavasti ja niiden hyödyntäminen data-analytiikassa on yleistynyt merkittävästi (Winter & Davidson 2019). Kasvavat tietomäärät erilaisista lähteistä, kuten tietojärjestelmistä, sensoreista ja päätelaitteista aiheuttavat tietojen epäjohdonmukaisuuksia, kun tietoja kootaan yhteen. Tietojen epäjohdonmukaisuudet on tunnistettava ja korjattava ennen kuin päätöksiä tehdään sen pohjalta, jotta vältytään väärien päätöksien tekemiseltä virheellisen tiedon pohjalta. (Abraham ym. 2019.) Uudenlaiset data-analytiikkaratkaisut lupaavat tarjota yhä parempia terveyspalveluita kuluttajille (Winter & Davidson 2019). Dataan liittyen yksi suurimmista haasteista yrityksille on se, pystyykö se vastaamaan haluttuihin kysymyksiin, koska esimerkiksi ei voida tietää millaiset ennustemallit ovat sopivia datalle. Lisäksi on vaikea sanoa mitkä muuttujat pystyvät selittämään ilmiötä, josta yritykset haluavat saada syvempää tietoa. Data voi olla myös heikkolaatuista, mitä yritykset eivät välttämättä pysty hahmottamaan etukäteen. Tietojen heikko laatu johtaa analyysivaiheessa virheellisiin johtopäätöksiin ja tuloksiin. (Rahm & Do 2000.)

Organisaatioiden pitää siis yhä huolellisemmin kiinnittää huomiota tiedonhallintaan, kun tarkoituksena on tuottaa lisäarvoa yritykselle erilaisista tietoresursseista. Toimivaa tiedonhallintamallia suunniteltaessa on kiinnitettävä erityistä huomiota yrityksen ydinasioihin ja päätöksentekoon. (Khatri & Brown 2010.) Tiedonhallintamallin tarkoituksena on kertoa mitä päätöksiä tiedonhallinnan ja datan käytön varmistamiseksi tehdään ja mitä on tehtävä, jotta lisätään datasta hyödynnettävää arvoa sekä minimoidaan siihen liittyviä kustannuksia ja mahdollisia riskejä. Sen tavoitteena on myös toteuttaa koko yritystä kattava agenda, koskien dataa ja sen hyödyntämistä. Valtion instituutiot painottavat tiedonhallintamallin olemassaoloa nykyään huomattavasti enemmän. (Abraham ym. 2019; Khatri & Brown 2010.) Tiedonhallintamalli auttaa tietojoukkojen eheyden ja luotettavuuden varmistamisessa. Tämä on tärkeää dataprojekteja ajatellen, jotta aikaa ei kuluisi liikaa projektin alkuvaiheisiin, kuten projektin ymmärrykseen ja datan ymmärrykseen. CRISP-DM (englanniksi *CRoss Industry Standard Process for Data Mining)* on yleinen toimialasta riippumaton malli dataprojekteille. Se sopii dataprojekteille, joissa on tarkoitus hyödyntää data-analytiikkaa tai tekoälyä. CRISP-

DM:n eri vaiheita ovat liiketoiminnan ymmärrys, datan ymmärrys, datan valmistelu, mallinnus, mallin arviointi ja käyttöönotto. (Schröer ym. 2021.)

Kun yritykset ottavat käyttöön yhä enemmän analytiikkaa, tarvitaan koko organisaation kattava ymmärrys datasta. GDPR (englanniksi *General Data Protection Regulation*), eli Euroopan unionin yleinen tietosuoja-asetus, lisää yritysten painetta hallita dataa. (Abraham ym. 2019.) Data-analytiikkaratkaisut, jotka prosessoivat henkilökohtaisia terveystietoja kasvattavat huolta ihmisten yksityisyydensuojasta ja tietojen turvallisuudesta. Samalla kuitenkin ratkaisujen luvataan tarjoavan yhä parempia terveydenhuollon palveluja ja henkilökohtaisten terveystietojen käsittelystä on tehty helpompaa. (Winter & Davidson 2019.) Data-analytiikan ja tekoälyn hyödyntäminen luo mahdollisuuden ennaltaehkäiseviin sekä terveyttä edistäviin ratkaisuihin ja niissä on suuri markkinapotentiaali (Tevameri 2018). Yksilöiden terveystiedot ovat kuitenkin yhä useammin alttiina hyväksikäytölle ilman vastuuvelvollisuutta ja nämä terveystiedot ovat tulossa yhä helpommin yksityisten yritysten ja hallitusten saataville, eikä päinvastoin (Germann & Jasper 2020).

Terveystietojen keräämisessä ja käsittelyssä on runsaasti potentiaalia ja se tarjoaa erinomaisen mahdollisuuden terveysteknologia-alan yrityksille kehittää yhä parempia palveluja ja tuotteita terveydenhuoltopalvelujen tuottajille, mutta tiedonhallintamallit eivät ole kyenneet pysymään tämän kehityksen tahdissa. Tarpeeksi kattavien poliittisten toimenpiteiden, oikeudellisten ohjeistuksien ja sääntelykehysten puute on johtanut siihen, että nykyinen digitaalisten terveydenhuoltopalvelujen tutkimus ja kehitys tapahtuu pitkälti autonomisella kentällä, jossa hyödynnetään yksilöiden ja jopa kokonaisien yhteiskuntien terveystietoja. Hallintomekanismeja on muodostettava kansallisella ja kansainvälisellä tasolla, jotta yksilöiden ja yhteiskuntien eettiset huolenaiheet ja oikeudet saadaan tasapainoon terveysalan massadatan tarjoamien mahdollisuuksien kanssa. Yksityisyyden ja tietojen omistajuuden suojaaminen on monimutkainen haaste yksilöiden terveystietojen hallinnassa, eikä data-analytiikan hyötyjä saisi saavuttaa tietosuojan ja ihmisten yksityisyyden kustannuksella. (Germann & Jasper 2020.)

Tiedonhallintamallin vaikutus data-analytiikka hankkeiden ja projektien onnistumisen kannalta on siis merkittävä. On myös syytä kiinnittää huomiota riskeihin ja haasteisiin, jotka liittyvät data-analytiikan implementointiin terveysteknologiassa. Analytiikka ja tekoälysovellukset, joita on implementoitu terveydenhuollonkäyttöön, kuten

kuvantaminen ja diagnostiikka muuttavat merkittävästi potilaan ja kliinisen hoitotyöntekijän suhdetta. Kysymykseksi nousee se, miten potilaan tietoinen suostumus vaikuttaa tekoälyn ja analytiikan käyttöön hoitotilanteessa? Tämä on kriittinen kysymys, johon ei ole vielä kiinnitetty tarpeeksi huomiota, vaikka tietoinen suostumus on välitön haaste näiden teknologioiden implementoinnissa kliiniseen hoivaympäristöön. (Gerke ym. 2020.)

Terveysteknologia-alalla toimivat yritykset kohtaavat jatkuvasti haasteita liittyen säännösten ja regulaatioiden noudattamiseen. Uusien sääntelyjen ja vaatimusten vuoksi yrityksille voi olla epäselvää, millaista dataa saa prosessoida. Tämä voi puolestaan vaikeuttaa innovaatioiden kehittämistä. Jatkuva innovaatioiden kehittäminen on terveydenhuollon palvelujen parantamisen ja terveydenhuollon kustannusten pienentämisen edellytys. Tekoälyinnovaatioita on runsaasti, mutta poliittiset päättäjät vaativat sopivia menetelmiä, joilla voidaan arvioida mahdollista lisäarvoa terveydenhuollolle. (Annoni ym. 2018.) Tekoälyn jatkuvan kehityksen kannalta tarvitaan sääntelykehystä, jonka avulla kehitystyötä tehdään. Yhdysvaltalainen FDA (englanniksi *Food and Drug Administration*), Euroopan komissio ja monet Euroopan maat kehittävät käytäntöjä data-analytiikan ja tekoälyn kehityksen säätelemiseksi, mutta prosessi vie paljon aikaa. (Muehlematter ym. 2021.)

1.2 Tutkimusaukko ja tutkimuskysymykset

Tämä tutkimus keskittyy selvittämään tiedonhallintamallin tärkeyttä ja tarpeellisuutta, kun yritykset kehittävät data-analytiikkaratkaisuja terveydenhuollon ympäristöön. Tiedonhallintamalli on suunnattu pääasiassa suurten yritysten tarpeisiin, eikä malleja ole juurikaan suunnattu pk-yritysten hyödynnettäväksi. Yleinen oletus on se, että suurten yritysten tiedonhallintamallit voidaan pienentää suoraan pk-yrityksille sopivaksi. (Begg & Caira 2011.) Tiedonhallintamallien vaikutusta ja hyödyntämistä pk-yrityksissä on tutkittu melko vähän (Shah & AlSousi 2021). Terveydenhuollon organisaatiot vaativat yrityksiä todistamaan sääntelyiden ja regulaatioiden mukaisen toiminnan. Tiedonhallintamallin tulisi olla yleinen lähestymistapa, jotta tietojen hallinta olisi vastuullista ja että se sopii sekä asiakkaiden että organisaation tarpeisiin (Alhassan ym. 2019). Tutkimuksen tarkoitus on selvittää mitä pk-yrityksen tiedonhallintamallin tulee sisältää ja ottaa huomioon, kun asiakkaana ovat terveydenhuollon organisaatiot ja kun kehitetään data-analytiikka- ja tekoäly ratkaisuja terveydenhuollon organisaatioiden

käyttöön. Yleisesti pk-yritykset eivät koe, että tiedonhallintamalli toisi mitään konkreettista hyötyä. Tiedonhallintamalli määritellään ainoastaan sidosryhmien vaatimusten vuoksi. (Begg & Caira 2011.)

Työ regulaatioiden ja säännösten noudattamisen tekemiseksi on aikaa vievää ja yritykset kokevat sen raskaana. Vaikka GDPR on monin tavoin hyödyllinen kansalaisten oikeuksien kannalta sekä yrityksien eettisen toiminnan varmistamiseksi digitaalisessa liiketoiminnassa, etenkin pienillä ja keskisuurilla yrityksillä on vaikeuksia ymmärtää mitä vaatimustenmukaisuus tarkoittaa GDPR-kontekstissa ja miten sitä noudatetaan. Pk-yrityksillä on myös vaikeuksia noudattaa GDPR:n pitkiä ja lukuisia artiklan kohtia. Pk-yrityksille on vaikeaa hahmottaa sitä, mitä GDPR:n laadulliset näkökohdat tarkoittavat teknisten ratkaisujen kannalta. Ilman lakitekstin tuntemusta, GDPR sanaston hahmottaminen teknologian kontekstissa on teknisille henkilöille vaikeaa. (Sirur ym. 2018.)

Tutkimuksen aihetta lähestytään seuraavien tutkimuskysymysten avulla:

1. Mitä hyötyä tiedonhallintamallin toteuttamisesta on terveysteknologia-alalla toimiville pk-yrityksille?
2. Mitä osa-alueita tiedonhallintamallin pitää sisältää, kun pk-yritykset valmistautuvat käyttämään data-analytiikkaa terveystietojen käsittelyä varten?
3. Miten tiedonhallintamalli vaikuttaa terveysteknologia-alan data-analytiikkaprojektien toteutukseen?

Tutkimus pyrkii tarjoamaan näiden tutkimuskysymysten avulla tietoa siitä, mitä hyötyä tiedonhallintamallin käyttöönotosta ylipäätänsä on, miten tiedonhallintamalli tulisi toteuttaa terveysteknologia-alan pk-yrityksissä sekä siitä, miten tiedonhallintamallin olemassaolo vaikuttaa data-analytiikkaprojektien läpivientiin pk-yrityksissä. Tutkimukseen liittyvä teoriaosuus ja kirjallisuuskatsauksen perusteella muodostettu tiedonhallintamallin viitekehys käsitellään luvussa 2. Luvussa 3 tarkastellaan muodostetun viitekehyksen vaikutusta CRISP-DM-metodologian eri vaiheisiin.

Tutkimus toteutetaan kvalitatiivisena tutkimuksena. Teoriaosuuden kirjallisuuskatsauksessa pyritään selvittämään sopivat tiedonhallintamallin viitekehyksen osat juuri terveysteknologia-alan pk-yrityksille. Selvitetyt viitekehyksen osat toimivat

empiriaosuuden haastatteluiden teemoina. Haastatteluiden avulla pyritään saamaan näkemys teoriaosuudessa muodostetun tiedonhallintamallin viitekehyksien osien hyödyllisyydestä.

1.3 Rajaukset

Koska tiedonhallintamalli kattaa hyvin laajasti useita dataan liittyviä osa-alueita organisaatioissa, tutkimus on rajattu tarkastelemaan tiedonhallintamallin osia, jotka vaikuttavat dataprojekteihin terveysteknologia-alalla. Lisäksi pyritään selvittämään juuri niitä tiedonhallintamallin elementtejä, jotka koskettavat CRISP-DM-prosessin vaiheita sekä terveysdatan käsittelyyn liittyviä asioita. Kattava tutkimus tiedonhallintamallien vaikutuksesta dataprojekteihin olisi liian laaja yhden Pro Gradu -tutkielman aiheeksi. Koska tiedonhallintamallien viitekehyksiä on valtava määrä, pyritään tässä tutkimuksessa keskittymään syvemmin niihin viitekehyksiin, jotka soveltuvat terveystietojen tarpeisiin. Terveystiedoilla tarkoitetaan dataa, joka on henkilön fyysiseen tai psyykkiseen terveyteen liittyvää tietoa, joita terveyspalvelujen tarjoajat keräävät. Näitä voidaan kutsua myös potilastiedoiksi, joita ovat muun muassa nimi, henkilötunnus, kuvantamistutkimuksen tulokset ja hoitokertomukset. (Euroopan parlamentti ja neuvosto 2016.) Perustavanlaatuisia tiedonhallintamalleja esitellään lyhyemmin, mutta tämä auttaa hahmottamaan paremmin tiedonhallintamalleja, jotka on luotu juuri terveystietojen analytiikkaa ajatellen.

Koska tiedonhallintamalli tässä tutkimuksessa koskee erityisesti tietojoukkoja, jotka terveydenhuollon organisaatio luovuttaa terveysteknologia-alan pk-yrityksen käsiteltäväksi, koskee se terveystietojen toissijaista käyttöä. Terveystietojen ensisijaisella käytöllä tarkoitetaan tietojen hyödyntämistä potilaan hoitoon, esimerkiksi sairaalassa, jossa tiedot on alun perin kerätty. Terveystietojen toissijaisella käytöllä tarkoitetaan terveystietojen käsittelyä sairaaloiden ulkopuolella, esimerkiksi kaupallisessa tarkoituksessa teknologia-alan yrityksissä. (Safran ym. 2007.)

CRISP-DM:n osalta tutkimus rajataan niihin asioihin, joihin tiedonhallintamalli vaikuttaa. Tässä tutkimuksessa ei syvennytä, esimerkiksi datan visualisoinnin eri vaiheisiin ja työkaluihin eikä myöskään eri algoritmeihin tai algoritmien tarkkuuden validointimetodeihin, vaikka nämä ovatkin olennaisia asioita CRISP-DM-prosessissa.

GDPR:n osalta pyritään myös tutkimus rajaamaan terveystietojen käsittelyyn liittyviin säännöksiin. Tämä tarkoittaa sitä, että tutkimuksessa tullaan käsittelemään ainakin säännöksiä, jotka koskevat henkilötietojen, kuten nimen ja osoitteen ja erityisten henkilötietojen, kuten biometristen ja terveydellisten tietojen käsittelyä. Tutkimuksessa käsitellään GDPR:n määrittelemien roolien, kuten rekisterinpitäjän ja tietojenkäsittelijän velvollisuuksia. Kattavampi GDPR:n läpikäynti ei ole tarpeellista, koska se sisältää useita säännöksiä, jotka eivät ole relevantteja tämän tutkimuksen kannalta. GDPR:n myötä tämä tutkimus rajoittuu vain suomalaisiin yrityksiin, joiden tulisi tallentaa tietojoukkoja ainoastaan EU:n alueella.

2 Tiedonhallintamalli

Ennen kuin perehdytään tarkemmin tiedonhallintamalliin (englanniksi *data governance*), on määriteltävä ensin mitä tieto tai oikeastaan data tarkoittaa tässä yhteydessä. Perinteisellä datalla tarkoitetaan strukturoitua dataa. Se on yleensä kvantitatiivista ja hyvin organisoitua. Esimerkkejä strukturoidusta datasta ovat esimerkiksi päivämäärät, nimet, luottokorttinumerot, henkilötunnukset sekä puhelinnumerot. Strukturoidun datan hyötyjä ovat helppo käsiteltävyys sekä helppo pääsy. Esimerkiksi yrityksien toiminnanohjausjärjestelmät säilövät strukturoitua dataa tietojoukoissa. Strukturoimaton data kategorisoidaan yleensä kvalitatiiviseksi dataksi. Strukturoimatonta dataa hallitaan parhaiten tietokannoissa, jossa data ei seuraa mitään ennalta määrättyä rakennetta. (IBM 2021.) Esimerkki tällaisesta tietokannasta on esimerkiksi NoSQL, missä tietojoukot eivät vaadi taulukoita ja jossa tieto skaalautuu vaakatasossa (Tudorica & Bucur 2011). Strukturoimaton data voi olla esimerkiksi tekstiä erilaisista dokumenteista, kuvia, sosiaalisen median julkaisuja tai sensorien ja mittalaitteiden tuottamaa dataa (IBM 2021). Tietojoukkoja, jotka kootaan useasta eri lähteestä ja joissa data esiintyy monessa eri muodossa, on hankala hallita. Tällaista tietojoukkoa voidaan kutsua massadataksi (englanniksi *Big Data*). Sillä tarkoitetaan sellaisia massiivisia tietojoukkoja, jotka sisältävät monimutkaista strukturoitua sekä strukturoimatonta dataa. Big Dataan liittyen yritykset kohtaavat useita haasteita, joita ovat hallinta, tietoturvallisuus, analysointi, käsittely ja säilöminen. (Al-Badi ym. 2018.)

Yritykset käyttävät monimutkaisten tietojoukkojen hallintaan tiedonhallintamallia. Tiedonhallintamalli on pohjimmiltaan tietoresurssien hallintaan liittyvien vastuualueiden ja päätäntävallan jakamista. Tutkijat ja ammatinharjoittajat ovat tästä yleensä yhtä mieltä. (Khatri & Brown 2010; Otto 2011.) Sillä tarkoitetaan tietoresurssien hallintaa sekä valvontaa. Sen avulla pyritään luomaan koko organisaatiota kattava suunnitelma datan käyttöä varten, saamaan datavarannoista mahdollisimman paljon hyötyä organisaatiolle sekä minimoimaan dataan liittyvät riskit. Koska nykyään datan hyödyntämisestä on tullut tärkeämpää yhä useammalle yritykselle, painotetaan sen olemassaoloa yhä enemmän. (Abraham ym. 2019; Weber ym. 2009.) Tiedonhallintamallin tarkoitus on taata tietovarantojen vastuullinen käsittely, tietojoukkojen turvallisuuden kehittäminen, tietojen ja liiketoiminnan kannalta olennaisten toimintojen välisen johdonmukaisuuden parantaminen sekä laadukkaan tiedon tarjoaminen käyttöön. Tiedonhallintamalli auttaa

myös tietojen järjestelmällisessä yhdenmukaistamisessa ja tietojenkäsittelyyn liittyvien hyvien käytäntöjen luomisessa. (Kim & Cho 2018.)

2.1 Tiedonhallintamallit data-analytiikassa

Data-analytiikan avulla pyritään saamaan ymmärrystä valtavista tietomääristä, joita yritykset säilövät tietovarastoihinsa. Valtavista tietomääristä johtuen tavanomaiset tiedonhallintamallit eivät ole riittäviä. Tavanomaiset tiedonhallintamallit ovat sopivia, esimerkiksi strukturoidun tiedon hallintaan, joka on saatavilla helposti, esimerkiksi yrityksen toiminnanohjausjärjestelmästä. Monimutkaisemman Big Datan käytön vuoksi yritykset kohtaavat ongelmia dataan liittyen, kuten käsittely, hallinta, tietoturva, analysointi, tallentaminen, etsiminen, jakaminen sekä visualisointi. Tämän vuoksi tarvitaan tiedonhallintamalleja, joita voidaan kutsua myös Big Data -hallintamalleiksi. Tällaisella tiedonhallintamallilla pyritään hallitsemaan yritysten valtavia tietomääriä sekä hyödyntämään dataa yritysten päätöksenteossa data-analytiikkaratkaisujen avulla. Big Data -hallintamallia varten tietoja on valmisteltava ensiksi, jotta tiedot olisivat johdonmukaisia ja luotettavia ja myös siksi, että lopulta voidaan luottaa analytiikan tuottamiin tuloksiin. (Al-Badi ym. 2018.)

Yrityksien on hallittava Big Dataa kuin mitä tahansa muitakin tietoresursseja. Hallintamalli on tämän vuoksi kriittinen, jotta saataisiin data-analytiikkaprojekteista suurin mahdollinen hyöty. (Niemi 2011.) Big Data -tietojoukkojen hallinta on yleensä hajaantunut usean eri organisaation välille. Tämä vaikeuttaa tietojoukon hallintaa ja vahingoittaa tietojen eheyttä. Ilman tiedonhallintamallia, tällaisten tietojoukkojen valvonta ja vaatimustenmukaisuuden varmistaminen on liian riskialtista. Siksi tiedonhallintamallia käytetään datan laadun valvomiseen sekä siihen, että varmistetaan lakien ja eettisten periaatteiden noudattaminen dataa hyödyntäessä. Data-analytiikan hyödyntämisessä on riskinsä, jos tietojoukkojen eheydestä tai luotettavuudesta ei ole huolehdittu tarpeeksi. Väärien analyysien perusteella voidaan tehdä täysin harhaisia tai laittomia päätöksiä, jotka voivat johtaa taloudellisiin, oikeudellisiin ja sosiaalisiin riskeihin. Yhä digitaalisemmassa maailmassa, jossa hallitukset, yritykset ja kansalaiset keräävät tietoja ja eri tahot käsittelevät niitä eri algoritmeilla, virheiden määrä kertyy ja vastuullisuus tietojoukoista häviää vähitellen kokonaan. (Janssen ym. 2020.)

Tiedonhallintamalli on myös hyödyllinen analytiikan ja tekoälyn tulosten kannalta. Hyvä tiedonhallintamalli varmistaa hyvän pohjan analytiikan ja tekoälyn toteutukseen, jolloin

myös niiden tuottamia tuloksia voidaan pitää luotettavampana. Tiedonhallintamalli selkeyttää organisaation, ihmisten, datan ja teknologian väliset suhteet yrityksessä. Vakiintunut tiedonhallintamalli varmistaa, että organisaatio on valmis hyödyntämään analytiikkaa ja tekoälyä. (Okoro 2021.)

Okoron (2021) mukaan yhä useammat päätökset tehdään nykyään erilaisilla koneoppimismalleilla ja tekoäly ohjaa päätöksentekoa datan avulla. Se tarkoittaa, että oli algoritmien käyttötarkoitus mikä tahansa, tiedonhallintamallin intressinä on taata vastuulliset, läpinäkyvät ja oikeudenmukaiset algoritmit. Lakien ja säädösten noudattaminen on erityisen tärkeää, koska mikään yritys ei halua joutua julkisuuteen lain rikkomisesta tai analytiikkasovelluksien aiheuttamasta syrjinnästä. Lakien noudattamista voidaan tukea noudattamalla tiedonhallintamallin viitekehystä. Lakien ja säädösten noudattaminen on erityisen tärkeää terveysteknologia-alalla, koska tietojenkäsittely koskettaa yksilöiden henkilökohtaisia terveystietoja.

2.2 Terveystiedot ja tiedonhallintamalli

Tiedonhallintamalli tarjoaa terveydenhuoltoalalle arvokkaan lähtökohdan dataperusteisille projekteille, kuten tietojen varastointiin, terveystietojen laadun parantamiseen, terveystietojen louhintaan ja terveystietojen hyödyntämiseen data-analytiikan avulla. Lisäksi tiedonhallintamalli voi tehostaa liiketoiminnan ohjausta sekä strategisten päätösten tekemistä. Nämä soveltamisalat ovat arvokkaita, mutta niitä on tutkittu tähän mennessä lähinnä lähtökohtana mahdollisuuksille, mitä datan hyödyntämisen avulla voidaan saavuttaa. (Lind & Glas 2022.)

Henkilökohtaiset terveystiedot ulottuvat yhä enemmän terveydenhuollon ja apteekkien ulkopuolelle. Yksilöiden terveystietoja voidaan kerätä suoraan päivittäisten toimintojen perusteella erilaisista lähteistä, kuten aktiivisuuskelloista ja verkkoselaimen tiedoista. Näiden tietolähteiden määrä kasvaa jatkuvasti ja niitä louhitaan yritysten omiin tarpeisiin tai sitten myydään eteenpäin. Tällaiset terveystiedot jäävät helposti nykyisten terveystietosäännösten ulkopuolelle. Näitä tietoja säätelevät teknologiayritysten omat tietosuojakäytännöt ja näin ollen raja terveysdatan ja tavanomaisen asiakasdatan välillä on häilyvä. Data-analytiikan avulla luodut ennustavat mallit terveysdatasta voivat tehdä sellaisia päätöksiä, jotka eivät ole yksilöiden edun mukaisia. (Winter 2021.)

Henkilökohtaisen terveystietojen turvallisuus on valtion säädösten alaista, koska niiden käyttö voi aiheuttaa syrjintää ihmisiä tai ihmisryhmiä kohtaan terveydentilan perusteella. Tästä syystä terveystietojen tietoturva on suuri huolenaihe, kun pyritään estämään niiden luvaton käyttö. Nämä huolenaiheet ovat lisänneet tiedonhallintamallien huomiota terveystietoihin liittyen. (Winter & Davidson 2019.) Terveystietojen tulisi olla myös läpinäkyvää kaikille, joista tietoa on kerätty. Potilaat ovat alkaneet vaatia pääsyä omiin tietoihinsa, joka kerätään esimerkiksi kehoon kiinnitettyjen laitteiden avulla. (Hripcsak ym. 2014.) GDPR vaatii, että henkilölle, jonka tietoja käsitellään, on ilmoitettava kaikista tietojenkäsittely menetelmistä, joissa henkilön tietoja, kuten terveystietoja käsitellään sekä tietojen läpinäkyvyyttä. Ilmoituksessa on ilmoitettava tietojen prosessoija, esimerkiksi teknologiayritys, tietojen haltija eli rekisterinpitäjä sekä tietojen säilytysaika. Henkilölle on myös ilmoitettava automaattisesta päätöksenteosta tietoihin liittyen. Automaattinen päätöksenteko voi tarkoittaa, esimerkiksi data-analytiikka tai tekoälyratkaisuja. Ilmoituksessa on myös ilmettävä se, käytetäänkö kerättyä tietoa muihin tarkoituksiin kuin siihen mihin se on kerätty ja se, mitä jatkokäsittely sisältää. Data-perusteisissa järjestelmissä ilmoitusoikeuden tarkoitus on taata tietojen läpinäkyvyys yksilölle. (Kroll 2018.)

2.3 Tiedonhallintamallin hyödyt pk-yrityksille

Pk-yritykset kohtaavat edelleen kehityspaineita jatkuvasti kehittyvässä digitaalisessa liiketoimintaympäristössä, kun yrityksille asetetaan merkittäviä vaatimuksia tiedonhallintaan. Tämä voi johtaa siihen, että yritykset ottavat tiedonhallintamallin käyttöön vasten tahtoaan. (Begg & Caira 2011.) Tiedonhallintamallin hyödyntäminen auttaa pk-yrityksiä olemaan vaatimustenmukaisia tietoturvan ja tietosuojan suhteen. Asiakkaat vaativat, esimerkiksi GDPR:n noudattamista eikä sopimuksia tai kauppaa tehdä ennen kuin yritykset ovat todistaneet vaatimustenmukaisen tietosuojatason.

Pk-yrityksillä tulisi olla sisäänrakennettu tietosuoja, jossa pidetään huolta tietojen anonymisoinnista sekä henkilötietojen käsittelytoimista. Lisäksi tietoja tulee käsitellä vain ennalta määriteltyihin tarkoituksiin. (Okoro 2021.) Henkilötietojen anonymisointi tarkoittaa tietojen muokkaamista data-analytiikkaa varten niin, ettei henkilöä pysty tunnistamaan tiedon perusteella (Bäck & Keränen 2017). Useimmissa tapauksissa terveysteknologia-alalla pk-yritys on todellinen vastuullinen tietojenkäsittelijä ja varsinainen hyödynnettävä data saadaan, esimerkiksi terveydenhuollon organisaatiolta.

Tässä tapauksessa tietojenkäsittelijäyrityksien ei tarvitse huolehtia siitä, miten ja millä perustein terveystiedot on kerätty yksilöiltä ja onko keräämisessä noudatettu GDPR:n vaatimuksia. Tietojenkäsittelijäyrityksien on kuitenkin osoitettava olevansa vaatimustenmukaisia tietosuojaan liittyen, koska tietojenkäsittelijän on toimittava alkuperäisen rekisterinpitäjän määräyksien ja vaatimusten mukaan. Jos tietojenkäsittelyä suoritetaan rekisterinpitäjän puolesta, rekisterinpitäjän on hyväksyttävä ainoastaan sellaisia tietojenkäsittelijöitä, jotka toteuttavat riittävät suojatoimet teknisestä ja organistorisesta näkökulmasta. Tietojenkäsittelijän on täytettävä GDPR:n vaatimukset ja varmistettava yksilöiden oikeuksien suojelu liittyen henkilötietojen prosessointiin. (Euroopan parlamentti ja neuvosto 2016.) Terveysteknologia-alan yrityksille asetetaan julkisen hallinnon kilpailutuksessa jo vaatimukset tietosuojan toteuttamisesta. Jos toiminta ei ole regulaatioiden mukaista, asiakkaat eli terveydenhuollon organisaatiot eivät halua hankkia yrityksen ratkaisua. Tiedonhallintamallin luo standardit vaatimustenmukaisuuden toteuttamiselle, joten sen käyttöönotto on pk-yrityksien menestymisen kannalta terveysteknologia-alalla elintärkeää.

Tiedonhallintamalli on tärkeässä asemassa, jotta voidaan varmistaa tietojen oikeellisuus ja laatu data-analytiikkaprojektin jokaisessa vaiheessa. Tietojen oikeellisuus analyysissä, ennustamisessa ja muokkaamisessa estävät väärinkäytökset ja puolueelliset tulokset. Tiedonhallintamalli varmistaa tietoturvaan ja tietosuojaan liittyvät validointitarkastukset, jotta tietojoukkoja ei manipuloida tai käytetä väärin. Ensimmäinen tiedonhallintamallin ja data-analytiikan yhdistävä tekijä on Big Data, jota esiintyy lähes kaikkialla. Big Datan eheys saattaa kärsiä tietojoukkoja siirrettäessä ja useimmiten isot tietomurrot tapahtuvat juuri siirron aikana. Tietoturvan tulisi olla myös data-analytiikka- ja tekoälyprojekteihin liittyen etusijalla projektien alkuvaiheessa, varsinkin terveysteknologia-alalla. (Okoro 2021.)

Tiedonhallintamallissa määritellyt toimintatavat GDPR:n noudattamisen suhteen auttavat myös yrityksiä toteuttamaan eettisesti kestäviä data-analytiikka ja tekoälyratkaisuja. Yrityksien on oltava tietoisia terveystietojen etiikasta, jotta varmistutaan oikeudenmukaisesta datan käytöstä. Terveystietojen käsittely eettisesti rakentaa parempaa asiakasluottamusta ja se samalla osoittaa yrityksen arvokkuuden toimijana. Tietojen kestävä eettinen prosessointi vähentää myös tietoihin liittyvää puolueellisuutta, sosiaalista epäoikeudenmukaisuutta ja lisää säädöksien noudattamista (Okoro 2021.) Terveysteknologia-alan yrityksillä on suuri riski joutua näkyvän julkisen tarkastelun

kohteeksi, jos tiedonhallinta-asioita ei hoideta vaatimusten mukaisesti ja eettisesti oikein. Tämä johtuu terveysalan luonteesta ja yhteiskunnan normeista. Jos väärinkäytökset kohdistuvat heikommassa asemassa oleviin ihmisiin ja heidän henkilötietoihinsa ja ominaisuuksiin, haitat ovat niin merkittävä, että yritys tuskin pystyy jatkamaan toimintaansa. Ihmishengen menetys johtuen data-analytiikan tai tekoälyn virheestä on mahdollista ja se aiheuttaisi katastrofaaliset seuraukset yritykselle. Tämä johtaisi maineen, luottamuksen ja tulojen menetykseen, rikostutkintaan sekä lakien ja sääntelyjen antamiin sanktioihin. (Cheatham ym. 2019.)

2.4 Tiedonhallintamallien viitekehyksiä

Tiedonhallintamallien viitekehys tarjoaa merkittävää apua organisaatiolle, kun tarkoituksena on yhdistää organisatoriset ja tekniset monimutkaisuudet, kuten laaja tietojoukko organisaation käyttöön (Panian 2010). Tiedonhallintamalli määrittelee viitekehyksen tietojen hallintaan, joka voidaan nähdä myös yrityksen strategisena voimavarana. Tiedonhallinta määrittelee oikeudet päätöksenteolle ja vastuut dataan liittyvässä päätöksenteossa. Lisäksi tiedonhallintamalli selkeyttää dataan liittyvät käytännöt, standardit ja menettelyt sekä valvoo niiden noudattamista (Abraham ym. 2019.)

Khatrin ja Brownin (2010) mukaan tiedonhallintamallin (Taulukko 1.) viisi olennaista päätösaluetta ovat datan periaatteet, datan laatu, metadata eli ydintieto, dataan pääsy ja datan elinkaari.

Taulukko 1. Tiedonhallintamallin päätösalueet (Khatri & Brown 2010)

Päätösalueet	**Määritelmät**
Datan periaatteet	Datan käyttötarkoitus Datan hyödyntäminen resurssina Eri säädöksien vaikutus datan käyttöön liiketoiminnassa
Datan laatu	Datan laadun vaatimukset Keinot datan laadun hallintaan Datan laadun arviointi
Metadata	Tietojen määrittely ja mallinnus niin, että ne ovat tulkittavissa Metatietojen pitäminen ajan tasalla
Datan käyttöoikeus	Tiedonsaantiin liittyvien käytäntöjen selvitys Tietoturvallisuuden varmistaminen

Päätösalueet	Määritelmät
	Menettelyt koskien varmuuskopiointia ja tietojen palauttamista
Datan elinkaari	Tietojen inventointi Selvitys datan määrittelystä, tuottamisesta, säilyttämisestä ja poistamisesta Lainsäädännön vaikutus vaatimustenmukaisuusongelmiin liittyen tietojen säilyttämiseen ja arkistointiin

Al-badin ym. (2018) esittelevät tietojenhallintamallin viitekehyksen (Taulukko 2.) laajojen tietojoukkojen analytiikkaa varten. Tämä malli koostuu kahdeksasta osasta, joita ovat organisaatiorakenteen tunnistaminen, sidosryhmien valinta, Big datan laajuuden määrittäminen, säädökset ja standardit, optimointi ja laskenta, datan laadun mittaus ja seuranta, datan säilöminen sekä kommunikointi ja hallinta.

Taulukko 2. Tiedonhallintamallin osat (Al-badi ym. 2018)

Osa	Määritelmät
Organisaation rakenteen tunnistaminen	Tiedonhallintamallin sovittaminen organisaation näkemyksiin ja tavoitteisiin
Sidosryhmien valinta	Olennaisten sidosryhmien, kuten datatieteilijöiden, data-analyytikoiden, tietovastaavan ja johtoryhmän tunnistaminen
Big Datan laajuuden määrittäminen	Datan laajuuden ymmärtäminen, jotta varmistutaan käytettävien tekniikoiden riittävyydestä
Säädökset ja standardit	Sääntöjen ja standardien varmistaminen koskien tietojen keräämisestä, hallintaa, käyttöä, yksityisyyttä, turvallisuutta, riskejä ja luokittelua Varmistettava ovatko käytännöt järjestelmän kanssa yhteensopivia
Optimointi ja laskenta	Tiedon keruu ja analysointimenetelmät
Datan laadun mittaus ja seuranta	Data laadun mittaaminen ja seuranta on erityisen tärkeää Dataa on putsattava käyttökelpoiseen muotoon poistamalla epäjohdonmukaisia ja virheellisiä tietoja Datan laatua seurattava koko projektin ajan alusta loppuun saakka
Datan säilöminen	Tietojen tallennus suojattuun paikkaan, mutta tietojen oltava myös käytettävissä tarvittaessa
Kommunikointi ja datanhallinta	Tuloksien välittäminen asiakkaille

Janssenin ym. (2020) mukaan tiedonhallintamalli (Taulukko 3.) on perusta luotettavalle tekoälylle. He esittelevät viitekehyksen, joka koostuu 13 periaatteesta, joita ovat tietojen laatu ja puolueellisuus, muuttuvien tulosten tunnistaminen, tietojen laajuus, virheiden tunnistaminen, läpinäkyvyys, tietojen erottelu, yksilöiden tiedonhallinta, ydintietojen kerääminen, tietojen käyttöoikeus, hajautettu tietojen säilytys, tietovastaava, dataan liittyvän vastuun jakaminen ja datan käyttökelpoisuus. Näiden periaatteiden noudattaminen luo hyvän lähtökohdan data-analytiikan ja tekoälyn hyödyntämiselle.

Taulukko 3. Tiedonhallintamallin periaatteet (Janssen ym. 2020)

Periaate	**Määritelmät**
Tietojen laatu ja puolueellisuus	Tietojen laatu sekä mahdollinen juurtunut puolueellisuus pitää arvioida
Muuttuvien tuloksien tunnistaminen	Kun algoritmien tuottamat tulokset muuttuvat, tulokset on validoitava ja tarkastettava muutoksien syyt
Tietojen laajuus	Jaettavan tiedon määrä kannattaa minimoida vain jakamalla se, mikä on tarpeellista
Virheiden tunnistaminen	Ihmisiä voidaan rohkaista palkkion avulla tunnistamaan datasta virheitä ja raportoimaan niistä
Läpinäkyvyys	Ihmisille ja organisaatioille tulee ilmoittaa, jos heidän tietojaan jaetaan avoimuuden varmistamiseksi ja väärinkäytösten estämiseksi
Tietojen erottelu	Henkilökohtainen ja arkaluontoinen tieto tulee erotella muista tiedoista
Yksilöiden tiedonhallinta	Yksilöillä ja organisaatioilla tulisi olla mahdollisuus tunnistaa omien tietojensa oikeellisuus
Ydintietojen kerääminen	Tietoja tulisi kerätä alkuperäisestä lähteestä, jotta varmistutaan niiden oikeellisuudesta ja siitä, miten tietoja kerätään
Tietojen käyttöoikeus	Käyttöoikeus annettava vain niille tahoille, jotka sitä tarvitsevat
Hajautettu tietojen säilytys	Hajautetut järjestelmät ovat vähemmän haavoittuvia Tietojen yhdistely ilman lupaa ei ole helppoa
Tietovastaava	Tietovastaavia on hyvä määrittää, jotta tietoresurssien hallinnan vastuuvelvollisuus on virallista
Vastuun jakaminen	Vastuu datasta tulee jakaa niin, ettei kukaan tietty henkilö ei voi käyttää tietoja väärin

Periaate	Määritelmät
Käyttökelpoisuus	Data pitää tunnistaa arvokkaaksi resurssiksi, jota voidaan hyödyntää analytiikassa ja tekoälyssä

Hripcsakin ym. (2014) esittelevät yhdeksän periaatetta (Taulukko 4.) terveystietojen käsittelyyn. Sähköisten potilastietojärjestelmien myötä terveystietoja säilötään ja asetetaan saataville yhä helpommin. Kliinisestä näkökulmasta näitä tietoja käytetään potilaiden terveyttä ja hyvinvointia koskevien päätöksien tekemiseen. Periaatteita ovat tietojen käyttöoikeus, tietojen laatu, tietojen jakaminen, oikeudet ja vastuut, tietojen läpinäkyvyys, datan tuoma hyöty, tietovastaava, datan käyttöperiaatteet ja terveydenhuollon sidosryhmät.

Taulukko 4. Tiedonhallintamalli terveystietojen käsittelyyn (Hripcsak ym. 2014)

Periaate	Määritelmät
Tietojen käyttöoikeus	Dataan pääsy sekä käyttö pitäisi nähdä yhteisenä hyötynä kaikille Datan on oltava saatavilla ja käyttökelpoista asianmukaisiin tarkoituksiin tietoturva huomioiden
Tietojen laatu	Terveystietojen on oltava johdonmukaisia, vertailukelpoisia, ajankohtaisia ja luotettavia Käyttäjien on voitava seurata, missä määrin tiedot ovat saavuttaneet nämä ominaisuudet Datan konteksti ja alkuperä ovat tärkeitä käyttökelpoisuuden määrittämiseksi
Tietojen jakaminen	Siiloissa olevien terveystietojen integrointi ja jakaminen on välttämätöntä tiedon optimaalisen käytön kannalta
Oikeudet ja vastuut	Sidosryhmien (potilaat, perheet, palveluntarjoajat, tutkijat, maksajat ja organisaatiot) oikeudet ja velvollisuudet terveystietojen keräämisen ja käytön kannalta on ymmärrettävä ja kunnioitettava
Tietojen läpinäkyvyys	Tietojen käytön on oltava läpinäkyviä kaikille
Datan tuoma hyöty	Tietojen käytön hyötyjä on mietittävä suhteessa mahdollisiin riskeihin ja kustannuksiin
Tietovastaava	Tietovastaavien on osoitettava sitoutuminen ja ymmärrys terveystietojen hallinnasta Tietojen käyttö ja hyödyntäminen lakien ja määräysten mukaan
Tietojen käyttöperiaatteet	Käyttöperiaatteet ja käytännöt eivät saa olla niin tiukkoja, että ne rajoittavat tai estävät uusien teknologioiden käyttöä

Periaate	Määritelmät
Terveydenhuollon sidosryhmät	Terveydenhuollon sidosryhmien on selvitettävä uusien tietolähteiden käytön hyötyjä ja riskejä Tiedonkäyttöperiaatteiden päivitys tarvittaessa

Winterin ja Davidsonin (2019) mukaan terveystietojen tiedonhallintamalliin (Taulukko 5.) kuuluu viisi eri ulottuvuutta, joita ovat tiedon lähde, sidosryhmät, arvonluonti, hallinnan tavoite sekä hallinnan muoto. Ulottuvuudet sisältävät paljon määritelmiä ja joitakin niistä voidaan havaita jo aikaisemmin esitellyissä viitekehyksissä.

Taulukko 5. Tiedonhallintamallin ulottuvuudet (Winter & Davidson 2019)

Ulottuvuus	Määritelmät
Tiedon lähde	Jonkin organisaation keräämät tiedot, kuten henkilökohtainen terveyshistoria ja lääkereseptit Henkilökohtaisesti luotu tieto, esimerkiksi aktiivisuusdata (aktiivisuuskellot) ja kliininen data tutkimuksista Digitaalinen käyttäytymisdata, esimerkiksi hakuhistoria selaimessa, ostokset nettikaupoissa ja paikkatietojen ilmoittaminen verkkosovelluksiin
Sidosryhmät	Suorat sidosryhmät: Yksilö, perheenjäsenet, terveyspalvelujen tarjoajat, työnantajat Epäsuorat sidosryhmät: Valtio, terveydenhoitoalan tutkijat Terveydenhuoltojärjestelmä: Terveysteknologiayritykset, lääkefirmat, lääkinnällisten laitteiden valmistajat
Arvonluonti	Terveyden edistäminen data-analytiikan/tekoälyn avulla: Yksilön terveys, terveydenhuoltojärjestelmän tehokkuus, kansanterveys, tietojen kaupallistaminen
Hallinnan tavoite	Varmistetaan ja ylläpidetään, esimerkiksi luottamusta hallintoon, tietoturvaan, tietosuojaan, yksityisyyteen Helpotetaan tietoihin pääsyä, analytiikan hyödyntämistä ja tuetaan uusia innovaatioita
Hallinnan muoto	Lailliset periaatteet: tietosuoja Regulaatiot: GDPR Yritykset: Tietovastaava Teknologiat: Algoritmit, tietoturvatyökalut Standardit: Yhteistoimivuusprotokollat

Tiedonhallintamallien viitekehyksiä on siis useita ja ne sisältävät erilaisia termejä olennaisille periaatteille ja ulottuvuuksille. Viitekehyksien laajuus myös vaihtelee periaatteiden määrän perusteella. Tiedonhallintamallien periaatteiden, päätösalueiden ja osien ulottuvuuksien sisältö ei eroa hirveästi eri viitekehyksien välillä, oli kyseessä sitten viitekehys data-analytiikalle tai terveystietojen hyödyntämiseen. Merkittävimmät periaatteet ja päätösalueet, jotka nousevat esiin eri viitekehyksissä ja GDPR:n vaatimuksissa tässä vaiheessa, ovat esimerkiksi tietojen läpinäkyvyys, tietovastaavan määrittäminen. Data-analytiikan kannalta merkittävin osa-alue, mikä nousee viitekehyksissä esiin, on tietojen laatu. Seuraavassa luvussa perehdytään pk-yritykselle sopivan tiedonhallintamallin kriteereihin ja valitaan sopivimmat tiedonhallintamallin osat kriteerien perusteella.

2.5 Tiedonhallintamallin viitekehyksen muodostaminen pk-yritykselle

Pk-yrityksille tiedonhallintamallin viitekehyksen tulisi olla sellainen, joka on helposti ja edullisesti toteutettavissa, huomioi pk-yrityksen hallinnon ja rakenteen, huomioi suppeamman roolituksen (Nwabude ym. 2014). Pk-yrityksille helppous on tärkeää, oli kyse sitten käyttöön otettava viitekehys tai työkalu, koska pieni joukko ihmisiä tekee päätökset liittyen teknologisiin liiketoimintaratkaisuihin. Yritys todennäköisesti hylkää tiedonhallintamallin, jos se on liian monimutkainen. Lisäksi toteutuksen ja henkilökunnan koulutukseen liittyvien kustannusten tulisi olla alhaiset. Uusia innovaatioita otetaan nopeammin käyttöön niiden ollessa edullisia. (Okoro 2021.)

Oletettavasti terveysteknologia-alan pk-yrityksissä on kuitenkin tarpeeksi teknistä osaamista. Lisäksi jos halutaan hyödyntää tiedonhallintamallia nimenomaan dataprojekteja varten, vähintään tekninen ymmärrys siitä mitä tehdään, on oltava. Tarvittaessa analytiikka ja tekoälyosaaminen voidaan ulkoistaa toiselle yritykselle tai taholle. Tiedonhallintamallin helppo toteutettavuus ja käyttöönotto on yksi keskeisistä kriteereistä. Tiedonhallintamallin viitekehykseen pyritään valitsemaan mahdollisimman vähän elementtejä, mutta kuitenkin niin, että valitut elementit perustuisivat regulaatioiden vaatimuksiin ja kirjallisuuskatsauksen viitekehyksiin. Viitekehyksen tulisi olla yksinkertainen ja helppo toteuttaa, jotta ponnistelu sen käyttöönottoon olisi minimaalinen (Nwabude ym. 2014; Okoro 2021). Khatrin ja Brownin (2010) tiedonhallintamallin viitekehyksen voidaan katsoa olevan yksinkertaisin ja helposti skaalautuvin kaiken kokoisten yritysten tarpeisiin. Tätä viitekehystä on kuitenkin testattu vain suurissa

yrityksissä, joten todisteet sen sopivuudesta pk-yrityksille ovat puutteelliset. (Begg & Caira 2012.)

Ottaen huomioon edellisessä kappaleessa esitellyt tiedonhallintamallit sekä GDPR, muodostetaan terveysteknologia-alan pk-yrityksille sopiva tiedonhallintamallin viitekehys (Taulukko 6.)

Taulukko 6. Ehdotettu tiedonhallintamallin viitekehys tiedonhallintaan terveysteknologiayrityksille

Osa	**Määritelmät**	**Viitteet**
Sidosryhmien vaatimukset	Yrityksien ja terveydenhuollon organisaatioiden oikeudet ja velvollisuudet terveystietojen keräämisen ja käytön kannalta on ymmärrettävä. Regulaatioita ja lakeja, kuten GDPR ja tietosuojalaki tulee noudattaa.	(Euroopan parlamentti ja neuvosto 2016; Hripcsak ym. 2014; Khatri & Brown 2010)
Tietojen läpinäkyvyys	Ihmisille ja organisaatioille tulee ilmoittaa, jos heidän tietojaan jaetaan ja käytetään avoimuuden varmistamiseksi ja väärinkäytösten estämiseksi.	(Euroopan parlamentti ja neuvosto 2016; Hripcsak ym. 2014; Janssen ym. 2020; Khatri & Brown 2010)
Tietojen laatu	Dataa on saatettava käyttökelpoiseen muotoon poistamalla epäjohdonmukaisia, virheellisiä ja puolueellisia tietoja. Datan oltava vertailukelpoista, ajankohtaista ja saatavilla, jotta voidaan tuottaa luotettavia analytiikka ja tekoälyratkaisuja. Datan laatua seurattava koko projektin ajan.	(Al-badi ym. 2018; Hripcsak ym. 2014; Janssen ym. 2020; Khatri & Brown 2010)
Tietojen laajuus	Eri tietolähteiden selvitys ja yhdistäminen, jotta tietojoukot eivät ole puutteellisia ja varmistetaan käytettävien prosessointi tapojen riittävyys. Jaetaan vain tarpeellinen tieto.	(Al-badi ym. 2018; Euroopan parlamentti ja neuvosto 2016; Janssen ym. 2020)
Tietovastaavan määrittäminen	Yrityksien on määriteltävä tietovastaava, jotta tietojen hallinnan vastuuvelvollisuus on virallista. Terveydenhuollon sidosryhmät ja GDPR	(Euroopan parlamentti ja neuvosto 2016; Hripcsak ym. 2014; Janssen ym. 2020; Winter & Davidson 2019)

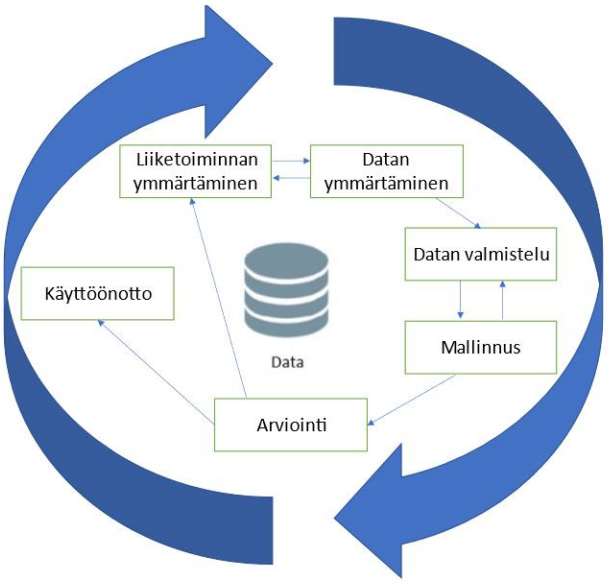
Osa	Määritelmät	Viitteet
	vaativat tietosuojavastaavan määrittämistä.	
Tietojen prosessointi	Datan prosessointimetodit on määriteltävä niin, että ne ovat reiluja ja asianmukaisia yksilön suhteen. Arkaluontoisten henkilötietojen anonymisointi on varmistettava.	(Al-badi ym. 2018; Euroopan parlamentti ja neuvosto 2016; Winter & Davidson 2019)

Ehdotetusta viitekehyksestä on pyritty tekemään mahdollisimman yksinkertainen terveysteknologia-alan pk-yrityksen tarpeisiin ja siinä on myös pyritty huomioimaan terveysteknologia-alan olennaiset vaatimukset. Terveystietojen eettinen ja kestävä käsittely on myös huomioitu viitekehyksen muodostamisessa tietojen läpinäkyvyys- ja prosessointiosissa. Tietojen prosessoinnin on noudatettava eettisiä periaatteita, eikä yksilöiden turvallisuus saa vaarantua prosessoinnin seurauksena (Okoro 2021).

3 CRISP-DM

Isojen tietojoukkojen käsittely päätöksenteon tueksi on saanut yhä enemmän huomiota yrityksien IT-strategioissa. Dataprojekteissa voidaan hyödyntää yhä enemmän analyyttisia malleja, mutta dataprojektitiimeille on yleensä vaikeaa noudattaa tietynlaista projektimetodologiaa. Nykyisissä data-analytiikkaprojekteissa prosessimallit eivät ole vakiintuneita, vaikka erilaisia prosessimalleja on ollut useita vuosia saatavilla. (Schröer ym. 2021.) Vain 18 % data-analytiikkaprojektien tiimeistä noudattaa täsmällisesti jotakin prosessimallia (Saltz ym. 2018).

CRISP-DM on yleisesti tunnettu, yleisimmin käytetty sekä toimialasta riippumaton standardi dataprojekteille. CRISP-DM sisältää kuusi iteratiivista vaihetta, joita ovat liiketoiminnan ymmärtäminen, datan ymmärtäminen, datan valmistelu, mallinnus, arviointi ja käyttöönotto (Kuvio 1.) (Berthold ym. 2010; Huber ym. 2019; Schröer ym. 2021). Seuraavissa luvuissa perehdytään jokaiseen prosessimallin vaiheeseen ja pohditaan terveysdatan käsittelyyn sopivan tiedonhallintamallin yhteyttä kuhunkin vaiheeseen.



Kuvio 1. CRISP-DM -prosessimalli (mukaillen Schröer ym. 2021)

3.1 Liiketoiminnan ymmärtäminen

Dataprojekteissa käytetään suhteellisen vähän aikaa yrityksen liiketoiminnan ja projektin ymmärrykseen sekä datan ymmärrykseen verrattuna myöhempiin vaiheisiin. Näiden alkuvaiheiden merkitys on kuitenkin projektin onnistumisen kannalta huomattavasti merkittävämpi kuin esimerkiksi datan valmistelu tai mallinnus. Bertholdin ym. (2010, 26) mukaan: ”Vaikka projekteihin ja datan ymmärtämiseen käytetty aika on pieni verrattuna tietojen valmisteluun ja mallintamiseen (20 % : 80 %), projektin menestymisen kannalta merkitys on juuri päinvastainen (80 % : 20 %).”

Liiketoiminnan ymmärrysvaihe sisältää paljon kokouksia ja tarpeellisten dokumenttien läpikäymistä. Tärkeimmät asiat dataprojektin kohteesta tulee selvittää tässä vaiheessa. (Rodrigues 2020.) Yleisesti dataprojekteihin liittyy ongelmia, niin analyytikon kuin projektin omistajan näkökulmasta. Riskit liittyvät puutteelliseen kommunikaatioon, ymmärryksen puutteeseen ja epäselvään organisaatioon. Esimerkiksi pk-yrityksen data-analyytikolla voi olla vaikeuksia ymmärtää se, mistä projektissa on pohjimmiltaan kyse ja hahmottaa projektin sidosryhmiä ja vaatimuksia. Projektin omistajalla voi olla vaikeuksia ymmärtää dataan liittyviä vaatimuksia ja näitä vaatimuksia voi joutua hyväksymään myöhemmin, kun datassa ilmenee ongelmia. (Berthold ym. 2010, 26.) Koska terveysteknologia-alalla asiakkaat ovat pääasiassa terveydenhuollon organisaatioita, regulaatioiden mukaiset vaatimukset datalle pitäisi aina olla tiedossa. Tässä vaiheessa relevantteja tiedonhallintamallin osia ovat sidosryhmien vaatimukset, tietojen läpinäkyvyys, tietojen laajuus sekä tietovastaavan määrittäminen.

Sidosryhmien vaatimukset perustuvat pitkälti tietosuojalakiin ja GDPR:n sanelemiin säädöksiin tietojen prosessoinnista. Terveysteknologia-alan pk-yrityksien ja terveydenhuollon organisaatioiden on noudatettava näitä säännöksiä. (Euroopan parlamentti ja neuvosto 2016; Hripcsak ym. 2014; Khatri Brown 2010.) Tietojen prosessointia varten GDPR asettaa kuusi vaatimusta, joita yrityksien ja terveydenhuollon organisaatioiden on noudatettava. Ensimmäinen vaatimus liittyy tietojen prosessoinnin laillisuuteen, läpinäkyvyyteen ja reiluuteen. Datan käsittelyn on oltava reilua ja oikeassa suhteessa yksilöön nähden. Toinen vaatimus koskee käyttötarkoituksen rajoittamista. Tietojen prosessointi tulee suorittaa vain määriteltyyn tarkoitukseen eikä mihinkään muuhun. Kolmas vaatimus liittyy tietojen minimointiin, jolla tarkoitetaan vain olennaisen tiedon keräämistä. Neljäs vaatimus on tietojen tarkkuus. Tämän vaatimuksen mukaan

tietojoukkoja tulee päivittää heti, jos havaitaan epätarkkuutta tiedoissa. Viides vaatimus on henkilötietojen poistaminen heti, jos niitä ei tarvita enää. Kuudes vaatimus koskee tietoturvaa. Pk-yrityksien on todistettava tietoturvan toteutuminen teknisesti ja hallinnollisesti tietojen prosessoinnin aikana. (Euroopan parlamentti ja neuvosto 2016.) Tiedonhallintamallin sidosryhmien vaatimukset osan tarkoitus on siis saada selvyys näihin vaatimuksiin. Pk-yrityksien on vastattava suurimpaan osaan näistä vaatimuksista dokumentoinnin muodossa, jotta heillä on lupa prosessoida terveystietoja. Terveydenhuollon organisaatio rekisterinpitäjänä on vastuussa terveystietojen keräämisestä vaatimusten mukaan. (Euroopan parlamentti ja neuvosto 2016.)

Tietojen läpinäkyvyys pitää myös määritellä liiketoiminnan ymmärrysvaiheessa. Niille henkilöille, joiden tietoja käsitellään, on ilmoitettava tulevista prosessointitoimista. Näin varmistetaan avoimuus ja vältetään tietojen väärinkäytökset. (Hripcsak ym. 2014; Janssen ym. 2020; Khatri & Brown 2010.) Henkilölle, jonka terveystiedot kerätään, on oltava selvää millä tavalla hänen tietojansa kerätään ja miten ne prosessoidaan (Euroopan parlamentti ja neuvosto 2016). Samalla tavalla kuin tiedoista voidaan tehdä läpinäkyviä, myös data-analytiikassa ja tekoälyssä käytetyistä algoritmeista voidaan tehdä läpinäkyviä. Algoritmien läpinäkyvyys ei välttämättä hyödytä yksilöä, koska niitä on vaikea tulkita ilman asiantuntemusta. Sen sijaan sidosryhmiin kuuluvat tarkastajat, asiantuntijat ja tutkijat voisivat tarkastaa algoritmien toiminnan, jotta vältytään puolueellisia ratkaisuja tekevistä algoritmeista. (Janssen ym. 2020.)

Tietojoukkojen laajuuden selvittäminen on tärkeää dataprojektia tekevälle pk-yritykselle, koska käytettävät resurssit ovat rajalliset. Jos tietojoukot osoittautuvat odotettua laajemmiksi tai puutteellisiksi, projektin eteneminen hidastuu. (Al-badi ym. 2018.) Rekisterinpitäjältä eli asiakkaalta onkin siis hyvä selvittää muut mahdolliset tietojoukot, joita voidaan tarvita projektin toteuttamiseen. Asiakkaan vastuulla on pääasiassa noudattaa GDPR:n vaatimuksia, joka tarkoittaa tietojen laajuuden suhteen sitä, että vain ne tiedot on kerätty, joita prosessoidaan. (Euroopan parlamentti ja neuvosto 2016.)

Tietovastaavalla tarkoitetaan yrityksessä nimettyä henkilöä, joka vastaa yrityksen tietojen hallinnasta. Tietovastaava auttaa tiedonhallinnan ja ratkaisujen kehittämisestä. (Hovi 2020.) Yrityksien on määriteltävä tietovastaava täyttääkseen asiakkaan ja muiden sidosryhmien vaatimukset (Euroopan parlamentti ja neuvosto 2016). Tiedonhallintamallin perustana on vastuullinen tiedon kerääminen. Henkilökohtaisiin

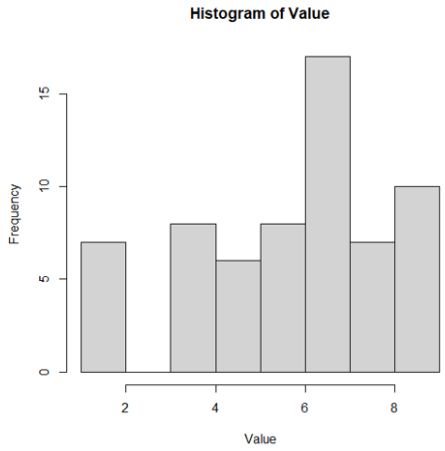
terveystietoihin voi kohdistua väärinkäytöksiä, joka aiheuttaa haittaa yksilöille. Tästä huolimatta useissa tilanteissa arkaluontoista tietoa on kerättävä yksilöiltä läpinäkyvyyden, palvelujen parantamisen ja parempien päätösten vuoksi. Tietoihin liittyvä vastuu on määriteltävä tämän vuoksi. Tietojen omistajuuden määrittäminen on usein haastavaa ja useat henkilöt voivat vaatia oikeuksia siihen. Yksilön edun tulisi kuitenkin olla etusijalla arkaluontoisten tietojen omistajuuden suhteen. Yrityksen nimeämän tietovastaavan tulee varmistaa vastuullinen tiedonjako sekä tiedonhallinta dataprojektin aikana. Tietovastaava on usein vastuussa tietojen laadun valvomisesta, tietoihin liittyvien riskien hallinnasta ja tietoturvan varmistamisesta. Tietovastaavan nimeäminen dataprojekteihin virallistaa myös vastuun tietojen hallinnasta yksilöiden edun mukaisesti. (Janssen ym. 2020.) Vastuun virallistamista vaativat myös dataprojektien asiakkaat, varsinkin kun kyseessä ovat yksilöiden terveystiedot.

3.2 Datan ymmärtäminen

Datan ymmärtämisvaiheessa pyritään selvittämään tietojen käyttökelpoisuus projektin tarkoitukseen (Rodrigues 2020). Erityisen tärkeää tässä vaiheessa on kartoittaa tietojoukoista sellaisia näkemyksiä, jotka ovat tärkeitä analyysin myöhemmissä vaiheissa. Lopullinen ratkaisu ei saa kuitenkaan ohjata datan ymmärrystä liikaa, muuten dataprojektin lopputulos voi olla puolueellinen. (Berthold ym. 2010, 33.) Tiedonhallintamallin avulla saadaan hyvät lähtökohdat dataprojekteille ja tietojen laadun parantamiselle (Lind & Glas 2022). Datan analysoinnin tulokset ovat täysin riippuvaisia datan laadusta ja laatu viittaa siihen, miten hyvin se sopii suunniteltuun käyttötarkoitukseen. Datan laadusta tarkastetaan tarkkuus, eheys, puolueellisuus ja ajankohtaisuus. (Berthold ym. 2010, 37-39; Hripcsak ym. 2014; Janssen ym. 2020.)

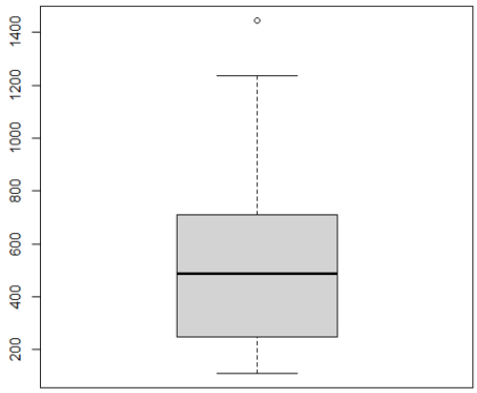
Tietojen tarkkuudella tarkoitetaan tiedoissa olevan arvon ja todellisen arvon välistä läheisyyttä. Numeeriselle datalle tarkkuus tarkoittaa tietojoukossa olevan arvon tarkkuutta suhteessa todelliseen arvoon. Rajallinen mittaustarkkuus huonontaa numeerisen datan tarkkuutta. Numeerisen datan tarkkuuteen voivat vaikuttaa myös manuaalisesti tallennetut arvot. Manuaalisesti tallentaessa tietoja tapahtuu helposti huolimattomuusvirheitä, esimerkiksi numeerista arvoa syöttäessä merkataan vahingossa yksi numero liikaa. Kategorisen datan eli tekstidatan tarkkuuteen voivat vaikuttaa, esimerkiksi kirjoitusvirheet. (Berthold ym. 2010, 37.) Tietojoukko on hyvä visualisoida, jotta epätarkat tai puuttuvat arvot voidaan tunnistaa ja korjata. Puuttuvia arvoja voidaan

tunnistaa, esimerkiksi histogrammin (Kuvio 2.) avulla. Histogrammi on hyvin samankaltainen kuin pylväskuvaaja, joka kuvaa jatkuvien muuttujien jakaumaa. Muuttuja jaetaan ensin sopiviin luokkiin. Luokkien havaintojen lukumäärää kuvataan sitten pylväällä, joka osoittaa havaintojen lukumäärän. (Nummenmaa ym. 2016, 50.) Histogrammista voidaan päätellä yhden palkin puuttumisen perusteella, että joitakin arvoja saattaa puuttua datasta ja asiaa pitää selvittää tarkemmin.



Kuvio 2. Histogrammi

Epätarkat arvot voidaan tunnistaa, esimerkiksi laatikkojanakuvion (Kuvio 3.) avulla. Laatikkojanakuviossa laatikon sisällä oleva kuvastaa mediaania. Laatikon alareuna kuvastaa alaneljännestä ja yläreuna yläneljännestä. Janan päät kuvaavat tietojoukon pienintä ja suurinta arvoa. Laatikon ja janan ulkopuolella olevat pisteet kuvastavat poikkeavia arvoja, jotka saattavat olla virheellisesti merkattuja. Poikkeavat arvot tulisi tarkistaa ja myös tarvittaessa selvittää niiden muodostumisen juurisyy, jotta niiltä voidaan välttyä jatkossa. (Taanila 2019.)



Kuvio 3. Laatikkojanakuvio

Tietojen eheys voi liittyä yksittäisiin puuttuviin arvoihin tai puutteelliseen kokonaiseen tietojoukkoon. Yksittäiset puuttuvat arvot on yleensä merkattu tietojoukkoon, jolloin puuttuvien arvojen mitta on niiden suhde muihin arvoihin. Puuttuvia yksittäisiä arvoja ei kuitenkaan välttämättä pysty tunnistamaan suoraan, joten niiden todellinen määrä voi olla suurempi kuin data antaa ymmärtää. Myös kokonaiset tietojoukot voivat olla puutteellisia. Joitakin olennaisia tietoja on hävinnyt, esimerkiksi sen vuoksi että eri tietolähteitä on yhdistelty ja joitakin tietoja jätettiin pois, koska koettiin että ne eivät ole enää tarpeellisia. (Berthold ym. 2010, 38.)

Dataa on lähestyttävä täysin neutraalista näkökulmasta, muuten dataprojektin lopputulos saattaa olla puolueellinen (Berthold ym. 2010, 33). Hyvin usein saatavilla oleva tietojoukko on puolueellinen tai se ei ole sopiva dataprojektin lopputuloksen kannalta (Berthold ym. 2010, 39). Tekoälyalgoritmeja opetetaan usein tietojoukoilla, jotka eivät edusta koko populaatiota. Kaushalin ym. 2020 mukaan: ”Esimerkiksi ihosyöpää havaitsevat algoritmit on usein opetettu datalla, joka on kerätty pitkälti vaaleaihoisilta. Nämä algoritmit toimivat huonommin, kun pyritään tunnistamaan ihosyöpää tummemmasta ihosta.”

Tietojoukon tulisi olla ajankohtainen käyttötarkoitustaan varten (Khatri & Brown 2010). Vanha data ei edusta tietojoukkoa eivätkä vanhalla datalla opetetut algoritmit luo hyödyllisiä ennusteita tulevasta. Tietojen ajankohtaisuus on vaikea säilyttää sellaisilla aloilla, joissa tieto on olennaista vain, jos se on kerätty äskettäin ja joissa vallitsevat trendit muuttuvat nopeasti. (Berthold ym. 2010, 39.)

3.3 Datan valmistelu

Datan ymmärtämisvaihe on edellytys datan valmistelulle. Tietojen ymmärtäminen auttaa tunnistamaan poikkeavia ja muuttuvia arvoja. Niiden käsittelyyn liittyvät, esimerkiksi arvojen muokkaaminen tai jättäminen tietojoukkoon sellaisenaan, ovat datan valmisteluun kuuluvia toimenpiteitä. (Berthold ym. 2010, 34.) Tiedonhallintamallin puute yrityksissä aiheuttaa sen, että tietoja ei ole standardoitu. Tämä tarkoittaa siis sitä, että tietojoukot ovat sekalaisia ja data ei ole järjestyksessä. Tämä aiheuttaa lisätyötä analyytikoille ja insinööreille, kun heidän on standardoitava tietojoukot putsaamalla datassa olevia virheitä. (Rodrigues 2020.) Kun tiedonhallintamalliin määriteltyjä tiedon laadun periaatteita noudatetaan yrityksessä, tällaista ongelmaa ei pitäisi muodostua.

Yrityksien ja sidosryhmien on tehtävä yhteistyötä, jotta tiedonhallintamallin noudattaminen on tehokasta. (Janssen ym. 2020.)

Datan valmisteluvaiheeseen kuuluu olennaisesti tietojoukon valinta, jota pyritään hyödyntämään. Joissakin tapauksissa dataa eri tietojoukoista on runsaasti saatavilla. Ongelmaksi muodostuu usein se, että ei pystytä varmistamaan datan relevanttiutta projektiin nähden. Usein saattaa käydä niin, että datan volyymi päihittää päätöksenteossa sen ajankohtaisuuden. Datan ymmärrysvaiheessa karsitaan yleensä pois huonot tietojoukot sekä epäolennaiset datan muuttujat. Data saattaa tästä huolimatta sisältää vielä turhia muuttujia, jotka eivät ole analyysin ja projektin lopputuloksen kannalta tarpeellisia. (Berthold ym. 2010, 116.) Datan liian suuri volyymi valmisteluvaiheessa johtaa siihen, että datan prosessointimetodien tehot jäävät puutteellisiksi (Al-badi ym. 2018). Tehokas prosessointi riippuu tietojoukkojen ja datassa olevien muuttujien määrästä (Berthold ym. 2010, 116). Datan liian suureen volyymin voidaan vaikuttaa tiedonhallintamallin avulla niin, että tietojoukkojen monimutkaisuus ja lähteet on määritelty hyvin edellisessä vaiheessa, jolloin datan suuren volyymin ei pitäisi tulla yllätyksenä enää tässä vaiheessa. Jaettavan tiedon määrä tulisi pitää maltillisena, jotta jaetaan vain olennainen tieto (Euroopan parlamentti ja neuvosto 2016; Janssen ym. 2020). Dataa on siistittävä käyttökelpoiseen muotoon poistamalla vääriä ja turhia tietoja. Tämä koskee niin yksittäisiä arvoja kuin myös muuttujia. (Al-badi ym. 2018.)

3.4 Datan mallinnus

Mallinnusvaihe on dataprojektien ydin. Mallinnusvaiheen tarkoitus on tuottaa datasta analytiikan avulla tuloksia, jotka vastaavat dataprojektin tavoitteita. Vaikka tämä vaihe on usein maineikkain CRISP-DM-vaiheista, se on myös kaikkein nopein vaihe, jos edellä olevat vaiheet on tehty huolella. (Rodrigues 2020.) Usein mallinnusvaiheessa oletetaan, että sopiva analytiikkaratkaisu voidaan valita useasta eri mallivaihtoehdosta. Liiketoiminnan ymmärtämisvaiheessa tehdyn työn pitäisi kuitenkin jo rajata analytiikkaratkaisujen joukkoa. (Berthold ym. 2010, 81.)

Tiedonhallintamallin tietojen prosessointi osa vaikuttaa datan mallinnusvaiheeseen. Hyvin suunniteltu, määritelty ja dokumentoitu tietojen prosessointi lisää tietoisuutta datan luonteesta ja käytettävistä algoritmeista ja se on yksi tiedonhallintamallin tehtävistä (Janssen ym. 2020). Varsinkin terveysteknologia-alan dataprojekteissa, analytiikkaratkaisun tai tekoälyn prosessointimetodin on oltava reilua ja

oikeudenmukaista yksilöön nähden. Yksilöiden tulisi päästä tarkastelemaan antamiaan terveystietoja kuten diagnooseja ja tutkimustuloksia, jotta he pysyvät perillä prosessoinnin vaatimustenmukaisuudesta. Yksilöiden yksityisyys tulisi myös varmistaa anonymisoimalla sellaisia tietoja, joista heidät voidaan tunnistaa. Koska kyse on terveystiedoista, yksilöiden tulisi halutessaan saada tietoa data-analytiikka- tai tekoälyratkaisusta sekä algoritmien toiminnan logiikasta. (Euroopan parlamentti ja neuvosto 2016.)

Tietojen laadulla on myös erittäin merkittävä vaikutus mallinnusvaiheeseen, koska algoritmien tuottamat tulokset riippuvat tietojen laadusta. Jos tietojen puolueellisuuteen ei ole puututtu tarpeeksi datan ymmärrysvaiheessa, algoritmien tuottamat tulokset ovat myös puolueellisia. Tekoälyalgoritmeja opetetaan usein sellaisella datalla, joka ei edusta koko populaatiota. Tämä johtaa tekoälyratkaisuihin, jotka eivät ole hyödyksi koko populaatiolle, vaikka se oli dataprojektin alkuperäinen tarkoitus. (Kaushal ym. 2020.)

3.5 Arviointi

Arviointivaiheessa yrityksen on varmistettava, että saadut tulokset vastaavat sidosryhmien odotuksia (Berthold ym. 2010, 297; Rodrigues 2020). Tiedonhallintamallin näkökulmasta, arviointivaiheessa voidaan myös arvioida tietojen laadun tilaa. CRISP-DM-prosessin aikana tietoja prosessoidaan yleensä useamman kerran, kun sitä sovitetaan valitulle mallille (Berthold ym. 2010, 298.) Datan laatu on varmistettava koko projektin ajan (Al-badi ym. 2018). Jos tulokset ovat oikeita, mutta eivät vastaa asiakkaan tarpeita, on palattava takaisin liiketoiminnan ymmärrysvaiheeseen ja selvitettävä, miksi arvioitava tulos ei vastannut dataprojektin tavoitteita (Rodrigues 2020).

Todisteet, havainnot ja johtopäätökset on dokumentoitava koko projektin ajalta tarkasti, jotta voidaan varmasti tehdä oikea päätös sen suhteen, otetaanko analytiikka tai tekoälyäratkaisua käyttöön seuraavassa vaiheessa. Tulokset, havainnot ja kaikki dataprojektin aikana suoritetut toimenpiteet tulee dokumentoida huolella. Dokumentoinnin merkitystä aliarvioidaan usein. Koska CRISP-DM ei ole lineaarinen prosessi, kaikki asiat mallin kehitykseen sekä tai huonontumiseen on dokumentoitava. Dokumentointi antaa edellytyksen paremman mallin jatkuvalle kehittämiselle. Muutokset mallin parantamiseksi saattavat sisältää uusien tietojen hyödyntämistä ja muokkaamista ennen mallinnusta. Kun muutoksia tehdään paljon mallin parantamiseksi, syntyy niin hyviä kuin huonoja ratkaisuja ja ilman dokumentointia kaikkia eri ratkaisuvaihtoehtoja

on vaikea hallita ja muistaa mitkä olivat hyviä ja mitkä huonoja. (Berthold ym. 2010, 298.)

Myös tiedonhallintamallin sidosryhmien vaatimusten ja tietojen läpinäkyvyyden kannalta dokumentointi on tärkeää. GDPR:n noudattaminen edellyttää sitä, että yksilö voi saada käsiin läpinäkyvästi kaikki hänen terveystietoihinsa liittyvät toimenpiteet. (Euroopan parlamentti ja neuvosto 2016.) Jos CRISP-DM:n eri vaiheissa dokumentointi on puutteellista, näitä yksilölle mahdollisesti tärkeitä tietoja ei pystytä tarjoamaan, mikä johtaa siihen, että niitä sidosryhmän vaatimuksia, jotka ovat olennaisia terveystietojen käsittelyyn liittyen, ei ole noudatettu.

3.6 Käyttöönotto

Viimeisessä käyttöönottovaiheessa tulokset esitetään asiakkaalle raportin muodossa. Raportti sisältää dataprojektin tulokset ja tärkeimmät havainnot. (Schröer ym. 2021.) Vaikka CRISP-DM-prosessin tarkoitus on luoda datasta syvempää tietämystä data-analytiikan ja tekoälyn avulla, ratkaisu on järjestettävä ja esitettävä niin, että asiakas ymmärtää mistä on kyse. Vaatimuksista riippuen, käyttöönottovaihe voi sisältää raportoinnin ja koko prosessin toistamisen. Käyttöönottovaihe korostaa sitä, että asiakas voi asettaa sellaisia vaatimuksia toimille, joilla varmistetaan arvon luonti CRISP-DM-prosessimallin avulla. (Lind & Glas 2022.) Asiakkaiden vaatimukset terveysteknologia-alalla ovat varmasti tiukimmasta päästä. Terveydenhuollon organisaatiot voivat vaatia tekoälyn toiminnan testaamista kliinisessä hoivaympäristössä ennen lopullista päätöstä. Tiedonhallintamallissa määritellyt sidosryhmien vaatimukset ovat siis suuressa osassa käyttöönottovaiheessa. Sidosryhmien vaatimukset sanelevat datan käytön periaatteet ja ne vaikuttavat yrityksen mahdollisuuksiin hyödyntää dataa haluamallaan tavalla (Khatri & Brown 2010).

3.7 Tiedonhallintamallin viitekehys ja CRISP-DM

Jokaisessa CRISP-DM-prosessin vaiheessa jokin tiedonhallintamallin osa vaikuttaa kyseisen prosessin onnistumiseen. Alla olevaan taulukkoon (Taulukko 7.) on tiivistetty tiedonhallintamallin osien vaikutus CRISP-DM-vaiheisiin. Merkittävin vaikutus tiedonhallintamallin viitekehyksellä on selkeästi CRISP-DM-prosessin alkuvaiheisiin, joita ovat liiketoiminnan ymmärtäminen ja data ymmärtäminen. Kun nämä alkuvaiheet ovat kunnossa, prosessin läpivienti onnistuu paremmin, eikä myöhemmissä, esimerkiksi

datan mallinnusvaiheessa ja arvioinnissa kohdata enää yllättäviä ongelmia, esimerkiksi tietosuojaan liittyen. Tiedonhallintamallissa määriteltyjen tietosuoja-asioiden tulisi olla etusijalla data-analytiikkaprojekteihin liittyen, varsinkin terveysteknologia-alalla (Okoro 2021). Liiketoiminnan ymmärtämiseen ja datan ymmärtämiseen käytetty aika on huomattavasti pienempi, mutta näiden vaiheiden merkitys projektin onnistumisen kannalta on todella suuri verrattuna, esimerkiksi datan valmisteluun ja mallintamiseen (Berthold ym. 2010, 26). Tiedonhallintamallin ja data-analytiikan ensimmäinen yhdistävä tekijä on Big data, eli laajat tietojoukot. Näiden tietojoukkojen eheys ja laatu saattaa kärsiä tietojoukkoja yhdisteltäessä. (Okoro 2021.) Myöhemmissä vaiheissa määrittelyt prosessointimetodit voivat olla puutteellisia, jos datan volyymi on liian suurta datan valmisteluvaiheessa (Al-badi ym. 2018).

Taulukko 7. Tiedonhallintamallin osien vaikutus CRISP-DM-vaiheisiin

Osa	**Määritelmät**	**CRISP-DM-vaihe**
Sidosryhmien vaatimukset	Yrityksien ja terveydenhuollon organisaatioiden oikeudet ja velvollisuudet terveystietojen keräämisen ja käytön kannalta on ymmärrettävä. Regulaatioita ja lakeja, kuten GDPR ja tietosuojalaki tulee noudattaa.	Liiketoiminnan ymmärtäminen, arviointi, käyttöönotto
Tietojen läpinäkyvyys	Ihmisille ja organisaatioille tulee ilmoittaa, jos heidän tietojaan jaetaan ja käytetään avoimuuden varmistamiseksi ja väärinkäytösten estämiseksi.	Liiketoiminnan ymmärtäminen, arviointi
Tietojen laatu	Dataa on saatettava käyttökelpoiseen muotoon poistamalla epäjohdonmukaisia, virheellisiä ja puolueellisia tietoja. Datan oltava vertailukelpoista, ajankohtaista ja saatavilla, jotta voidaan tuottaa luotettavia analytiikka ja tekoälyratkaisuja. Datan laatua seurattava koko projektin ajan.	Datan ymmärtäminen, datan valmistelu, datan mallinnus, arviointi
Tietojen laajuus	Eri tietolähteiden selvitys ja yhdistäminen, jotta tietojoukot eivät ole puutteellisia ja varmistetaan	Liiketoiminnan ymmärtäminen, datan ymmärtäminen

Osa	Määritelmät	CRISP-DM-vaihe
	käytettävien prosessointi tapojen riittävyys. Jaetaan vain tarpeellinen tieto.	
Tietovastaavan määrittäminen	Yrityksien on määriteltävä tietosuojavastaava, jotta tietojen hallinnan vastuuvelvollisuus on virallista. Terveydenhuollon sidosryhmät ja GDPR vaativat tietosuojavastaavan määrittämistä.	Liiketoiminnan ymmärtäminen
Tietojen prosessointi	Datan prosessointimetodit on määriteltävä niin, että ne ovat reiluja ja asianmukaisia yksilön suhteen. Arkaluontoisten henkilötietojen anonymisointi on varmistettava.	Datan mallinnus

4 Tutkimuksen toteutus

4.1 Metodologia

Tutkimusta lähestytään kvalitatiivisen tutkimusmenetelmän avulla. Kvalitatiivista tutkimusta käytetään yhä enemmän, kun arvioidaan jonkun teknologian vaikutusta, esimerkiksi yrityksen tietojärjestelmiin. Kvalitatiivisen tutkimuksen avulla pyritään ymmärtämään eri asioita tutkimalla ihmisten näkökulmia kontekstissa, jossa he toimivat. Tämän vuoksi kvalitatiivista tutkimusta tehdään kontekstin mukaisessa ympäristössä ja dataa kerätään numeroiden sijaan sanallisessa muodossa. (Kaplan & Maxwell 2005.)

Kvalitatiivinen tutkimus on erityisen merkityksellistä silloin, kun tarkasteltavan ilmiön aikaisemmat näkemykset ovat olleet vaatimattomia. Tämä voi johtua siitä, että tutkimusongelmaa on pyritty rajamaan strukturoiduilla rakenteilla, jolloin tulkinnat voivat jäädä vaatimattomiksi. Kvalitatiivinen tutkimus on joustavaa, jolloin se sopii hyvin strukturoimattomien ongelmien tarkasteluun. (Eriksson & Kovalainen 2008.) Kvalitatiivinen tutkimusmetodologia on sopiva tähän tutkimukseen, koska tutkittavaa ilmiötä pyritään ymmärtämään paremmin terveysteknologia-alalla toimivien pk-yrityksien näkökulmasta, mitä on tutkittu melko vähän.

4.2 Aineistonkeruumenetelmä

Laadullista tietoa kerätään ensisijaisesti haastatteluiden, havaintojen ja asiakirjojen avulla (Kaplan & Maxwell 2005). Tässä tutkimuksessa aineisto kerätään puolistrukturoiduilla teemahaastatteluilla. Teemahaastattelun avulla pyritään löytämään vastauksia tutkimuksen tarkoituksen ja tutkimuskysymysten mukaan. Teemat on valittu ennalta perustuen tutkimuksen viitekehykseen eli siihen mitä tutkimusaiheesta tiedetään jo entuudestaan. (Tuomi & Sarajärvi 2018.) Puolistrukturoidussa teemahaastattelussa haastattelijalla on valmis malli kysymyksistä ja teemoista, mutta haastattelija voi kuitenkin vaihdella kysymyksien järjestystä eri haastatteluissa. Etuna on se, että haastattelun materiaalit ovat systemaattisia, mutta haastattelun itsessään on pitkälti keskustelunomainen ja vapaa. (Eriksson & Kovalainen 2008.) Koska tiedonhallintamalli on moniulotteinen ja huomio monta yrityksen dataan liittyvää asiaa. Puolistrukturoitu teemahaastattelu on hyvä menetelmä kerätä haastateltavilta monipuolisia näkökulmia asiaan liittyen.

Koska tutkimuksen tarkoituksena on selvittää tiedonhallintamallin vaikutus dataprojekteihin terveysteknologia-alan pk-yrityksissä, haastateltaviksi valittiin sellaisia asiantuntijoita, jotka ovat olleet mukana tällaisissa dataprojekteissa tai dataprojektin valmisteluvaiheessa. Haastattelut toteutetaan etätapaamisina Zoomin avulla tai paikan päällä. Haastattelut tallennetaan haastateltavan suostumuksella, jotta kerättyä dataa pystytään analysoimaan huolellisesti. Tallenteet tuhotaan tutkimuksen valmistumisen jälkeen. Haastateltaville ilmoitetaan, että tutkimukseen osallistuminen on täysin vapaaehtoista ja että heistä ei kerätä mitään henkilötietoja aineistoa varten, koska se ei ole tutkimuksessa tarpeen. Tämän vuoksi tietosuojailmoitusta ei ole tarpeen tehdä. Haastateltavat anonymisoitiin tunnuksien avulla, esimerkiksi H1 tunnuksella. Haastateltaville lähetettiin myös heitä koskeva aineisto ja tulososio, jotta he varmistuivat tietojensa anonymisoinnista. Ennen haastatteluiden toteuttamista varmistettiin se, että haastateltavilla on kokemusta tiedonhallintamalleista terveysteknologia-alan pk-yrityksistä, joten tietoja haastateltavien tämänhetkisestä työnantajasta ei ole tarpeen kerätä aineistoon lainkaan. Haastateltavien rooli terveysteknologia-alan pk-yrityksissä varmistettiin myös, jotta kerätty aineisto olisi luotettavampi, ja että tulokset ovat uskottavampia. Alla olevassa taulukossa (Taulukko 8.) on esitelty tiedot haastateltavista. Haastateltaville informoitiin myös siitä, miten aineisto tallennetaan ja käsitellään tutkimuksessa aineistonhallintasuunnitelman (Liite 2.) mukaisesti. Aineistoa säilytetään Turun yliopiston suosituksen mukaan viisi vuotta tutkijan omalla tietokoneella, jossa on asianmukainen ja aina ajankohtaisesti päivitetty virustorjuntaohjelma.

Taulukko 8. Tiedot haastateltavista

Haastateltava	**Haastattelun kesto**	**Haastateltavan rooli**	**Kokemus terveysteknologia-alalta vuosina**
Haastattelu 1 (H1)	51 min	Koneoppimisinsinööri	1
Haastattelu 2 (H2)	58 min	Datatieteilijä	Lähes 30 vuoden kokemus data-analytiikasta sekä kokemus useammasta terveysdataprojektista
Haastattelu 3 (H3)	1 h 7 min	Terveysteknologiakonsultti	8
Haastattelu 4 (H4)	43 min	Tuotejohtaja	14
Haastattelu 5 (H5)	40 min	Tuotejohtaja	16

Toteutettava haastattelu jaetaan teoriaosuudessa määritellyn tiedonhallintamallin viitekehyksen mukaisiin teemoihin. Haastattelujen teemoja on yhteensä kuusi ja ne ovat sidosryhmien vaatimukset, tietojen läpinäkyvyys, tietojen laatu, tietojen laajuus, tietovastaavan määrittäminen ja tietojen prosessointi. Ennen kuin haastatteluissa siirrytään itse teemoihin, haastateltavilta kysytään muutama perustavanlaatuinen lämmittelykysymys. Grönforsin (2011) mukaan: ”Kun tutkijalla ei ole aikaisempaa suhdetta haastateltavan tai kun aikaisempi suhde on ollut vain satunnainen, on tutkijan tehtävänä luoda haastattelulle otollinen ilmapiiri. Tällaisissa tilanteissa on varsin tavallista aloittaa ns. verryttelykysymyksillä, joiden pääasiallisimpana tarkoituksena on myönteisen haastatteluilmapiirin luominen.” Haastattelukysymyksiä oli yhteensä 20, mukaan lukien johdannon lämmittelykysymykset. Jokaisessa haastattelussa hyödynnetään pitkälti samoja kysymyksiä keskustelun ylläpitämiseksi. Myös ennen jokaista haastattelun teemaa esitellään olennaiset termit ja se mihin nämä kysymykset liittyvät, jotta saadaan haluttua dataa tutkimuksen tarpeisiin. Haastattelun lopuksi on vielä osio vapaalle sanalle. Tämän avulla pyritään saamaan haastateltavalta vielä mielipiteitä tiedonhallintamallin viitekehyksestä yleisesti, esimerkiksi olisiko tästä viitekehyksestä hyötyä haastateltavan mielestä vai olisiko huomioitava vielä jokin muu asia tiedonhallintamallin osien lisäksi.

4.3 Analysointimenetelmät

Koska haastatteluja ohjaavat teemat perustuvat vahvasti kirjallisuuskatsauksessa luotuun tiedonhallintamallin viitekehykseen, kerätty aineisto analysoidaan hyödyntämällä teorialähtöistä analyysiä. Teorialähtöisessä analyysissä aineiston analysointi perustuu johonkin valmiiseen teoriaan tai malliin, joka on määritelty tutkimuksen teoriaosuudessa. Tätä valmista teoriaa tai mallia pyritään testaamaan analyysissä. (Tuomi & Sarajärvi 2018.) Tämä menetelmä sopii hyvin aineiston analysointiin, koska tarkoituksena on testata sitä, miten tiedonhallintamallin viitekehyksien osien sisältö on toteutunut ja vaikuttanut dataprojekteissa terveysteknologia-alan pk-yrityksissä.

Aineisto litteroidaan ennen analyysin aloittamista. Litterointi tehdään sanatarkasti Microsoft Word -tekstinkäsittelysovelluksen avulla. Seuraavaksi noudatetaan pitkälti Tuomen ja Sarajärven (2018) runkoa analyysin toteuttamisesta. Analyysin ensimmäisessä vaiheessa päätetään se, mikä aineistossa kiinnostaa ja mitä asioita poimitaan analyysiä varten. Toisessa vaiheessa kiinnostavat asiat merkataan aineistosta koodaamalla. (Tuomi

& Sarajärvi 2018.) Koodaus on hyvin yleinen tapa kvalitatiivisessa tutkimuksessa, jonka avulla tutkija jäsentelee aineistonsa ja pyrkii löytämään siitä jotakin uutta ja kiinnostavaa. Kun kiinnostavat asiat on kerätty, aineisto kootaan taas yhteen tutkimuksen kannalta mielekkäällä tavalla. (Elliot 2018.) Kiinnostavien asioiden koodaamiseen ei ole yleisohjetta. Tutkija voi itse päättää koodaukseen käytettävät merkit. (Tuomi & Sarajärvi 2018.) Koska haastatteluiden teemat noudattavat melko tarkkaan tiedonhallintamallin viitekehyksen osia, koodit perustuvat tarkemmin osien sisällä oleviin elementteihin. Tämän avulla valmistellaan myös viitekehyksen osien huolellista testausta, joka tehdään kolmannessa vaiheessa.

Tuomen & Sarajärven (2018) mukaan kolmannessa vaiheessa koodatut löydökset teemoitetaan ja tyypitellään. Tässä vaiheessa löydökset teemoitetaan haastatteluiden teemojen mukaan, mikä vastaa myös tiedonhallintamallin viitekehyksen osia. Tämän avulla korostetaan sitä, mitä kustakin teemasta on sanottu. Koodatut sitaatit kerätään teemojen mukaan taulukkoon. Aineisto myös tyypitellään eli teemojen sisältä etsitään yhteneväisiä asioita, jotka toistuvat ja joita voidaan pitää yleistyksinä. Viitekehyksen testauksen tulos saadaan aineiston tyypittelyn avulla hyvin selville. Neljännessä vaiheessa kirjoitetaan yhteenveto, jossa käydään läpi aineiston analyysin tulokset. (Tuomi & Sarajärvi 2018.)

5 Tulokset

Seuraavissa luvuissa 5.1–5.6 esitellään haastattelujen tulokset kuuden teeman mukaisesti. Lisäksi haastateltavien näkemyksiä esitetystä tiedonhallintamallista tarkastellaan luvussa 5.7. Havaintojen vaikutukset tiedonhallintamalliin esitetään luvussa 6.

5.1 Sidosryhmien vaatimukset

Haastatteluiden ensimmäinen teema koski sidosryhmien vaatimuksia. Tässä teemassa perehdyttiin myös suorien vaatimusten lisäksi myös asiakkaiden ymmärrykseen mahdollisista ratkaisuista. Heikko ymmärrys voi johtaa epärealistisiin odotuksiin koko projektista. Dataprojektien alkuvaiheissa esiintyy riskejä, jotka voivat liittyä kommunikaation epäselvyyteen ja ymmärryksen puutteeseen (Berthold ym. 2010, 26). Asiakkaiden teknisen ymmärryksen tasosta haastateltavilla oli poikkeavia mielipiteitä. Yleensä asiakkaalle tulee yllätyksenä se, mitä data-analytiikalla voidaan tehdä ja millaiset tekniset ratkaisut ovat mahdollisia. Tämä vaikutti siihen, että myös kommunikointiin oli panostettava enemmän. Kommunikointi koettiin selvänä, jos itse loppuratkaisu ei ollut kovin monimutkainen tai jos asiakas osaa hyvin kuvata sen mitä he haluavat algoritmin tekevän. Näissä tapauksissa myös itse pk-yritys oli erikoistunut melko kapeaan alueeseen tai pystyi tarjoamaan hyvin tarkkaan määriteltyyn tarkoitukseen ratkaisun. Kokemusta oli myös teknisen ymmärryksen parantumisesta, koska datan hyödyntämisen arvoa ymmärretään koko ajan paremmin.

> Koska kyseessä oli todella kapea käyttötarkoitus, missä käytettiin koneoppimismallia itse tuotteeseen, kommunikoinnissa ei ollut mitään epäselvyyksiä asiakkaan kanssa. (H1)

> Usein ollaan tekemisissä ihmisten kanssa, jotka ovat biologeja tai biologitaustaisia tai jotakin vähän sinnepäin kallellaan ja täytyy sanoa, että tekninen tietämys on aika heikko keskimäärin. (H2)

> Aika hyvin ymmärretään näissä projekteissa mitä asiakas haluaa. He osaavat kuvat sen mitä he haluaisivat algoritmin heille tekevän, ja me ymmärrämme mitä he haluavat, mutta he eivät tietenkään ymmärrä sitä miten tällaiset algoritmit kehitetään. (H4)

> Teknologian ymmärrys on hyvin vaihteleva, mutta parempi koko ajan, koska ymmärretään sen arvo. (H5)

Kommunikointi koettiin myös sellaisena asiana, johon on panostettava huomattavasti dataprojektien alkuvaiheessa, koska yhteinen terminologia ei ole selvää. Yhteisen

terminologian puute koettiin riskinä liian väljälle loppuratkaisun määrittelylle, joka voi johtaa siihen, että koko projektin tuotos jää puutteelliseksi asiakkaan kannalta. Kommunikointi koettiin myös haastavaksi, esimerkiksi tietoturvaan liittyen. Toki on ymmärrettävää, että terveysteknologia-alalla tietoturvaan kohdistetaan erityistä huomiota. Terveystietojen tietoturva on suuri huolenaihe, koska niiden väärinkäyttö saattaa aiheuttaa syrjintää ihmisiä tai ihmisryhmiä kohtaan (Winter & Davidson 2019).

> Kaikkein suurin ongelma on kyllä yleensä yhteisen terminologian ymmärtäminen, mitä tarkoitetaan milläkin asialla. Tämä on todella iso asia ja olen panostanut siihen monissa asiakkuuksissa. (H2)

> Kyllä hyvin yksinkertaisesti yritän selittää, että ei näihin tietoihin pääse kauhean helposti käsiksi, koska ne ovat sellaisten palveluiden takana, joka vaatii sitten ihan kunnon hakkerointia. Eikä ihan jokapojalla ole sellaiseen mahdollisuuksia. (H3)

Sidosryhmien vaatimukset tietojen käsittelyyn liittyen perustuivat pitkälti GDPR:n ja tietosuojalain vaatimuksiin tietojenkäsittelystä. Haastatteluissa ilmeni myös poikkeuksia tähän yleiseen käsitykseen. Näissä tapauksissa dataprojekti oli sidoksissa esimerkiksi hyvin harvinaiseen tietojoukkoon, jonka myötä sidosryhmän vaatimukset olivat hyvin tiukat. Tällainen johti lopulta siihen, että projekti ei päässyt alkua pidemmälle. Joissakin tapauksissa, vaikka asiakkaan tarjoama data oli anonymisoitu, eikä enää tietosuojan alaista, asiakkaat halusivat siitä huolimatta GDPR:n mukaisen selvityksen tietojenkäsittelymenetelmistä. Asiakkaat vaativat esimerkiksi salassapitolupausta sopimuksissa ja myös tietojenkäsittelysopimusta, jossa määritellään ehdot datan käsittelylle. Myös Aluehallintoviraston vaatimuksia pidettiin hankalina, koska vaatimukset vaihtelevat eri alueiden välillä.

> Yhtenä sidosryhmänä oli organisaatio, josta meidän piti saada dataa ja siinä itseasiassa kariutui koko homma siihen, että dataa ei saanut viedä mihinkään muuhun kuin sertifioituun ympäristöön organisaation vaatimusten mukaan. Suomessa oli sertifioituja ympäristöjä yksi ja siellä ei ollut tarjolla GPU laskentaa, mitä olisi tarvittu, joten ei voitu käyttää sitä dataa ollenkaan. (H2)

> Vaikka tämä data minkä he antavat onkin anonymisoitu ja tietosuojan puitteissa sen anonymisoidun datan ei pitäisi olla enää tietosuojan alaista dataa koska siinä ei enää henkilöitä voida tunnistaa, niin tästä huolimatta asiakkaat edelleen haluavat nähdä, että meidän datamme käsittely, tallennusmenetelmät ja muu hallintamalliin liittyvä on kunnossa. Esimerkiksi työntekijän sopimuksissa vaaditaan salassapitolupausta eli että ei saa jakaa näkemäänsä ja käsittelemäänsä tietoa ulospäin liittyen näihin datoihin. (H4)

> AVI:lla on säädöksiä ja erityyppisiä käytäntöjä, jotka vaihtelevat alueelta toiselle. Sellaisia on ihan oikeasti. (H5)

Yleisesti, haastateltavat kokivat GDPR:n vaatimuksien toteuttamisen erityisen työläänä pk-yrityksille, mutta kuitenkin positiivisena varsinkin yksilön kannalta, joista tietoa kerätään. GPDR on yksilöiden oikeuksien kannalta usealla tavalla hyödyllinen. Yritykset kuitenkin kokevat GDPR vaatimustenmukaisuuden toteuttamisen raskaana ja aikaa vievänä (Sirur ym. 2018.) Terveydenhuollon organisaation tietämystä GDPR:stä pidettiin hyvänä, mikä sinänsä ei ollut yllättävää, koska terveydenhuollon organisaatio on dataprojekteissa rekisterinpitäjän roolissa. Terveydenhuollon organisaation ja pk-yrityksien rooleja GDPR:n liittyen avattiin aiemmin luvussa 2.3.

> Omat kokemukset GDPR:stä on masentavan byrokraattisia, ymmärrän kyllä yksityisyyden suojan täysin. Mutta monessakin hyvä asia, etenkin terveysdataan liittyen aika tarpeen. (H2)

Muiden sidosryhmien tietämystä näistä vaatimuksista kuitenkin epäiltiin. Kaksi haastateltavista epäili melko voimakkaasti muiden sidosryhmien, kuten yhteistyökumppaneiden ja päättävien tahojen tietämystä GDPR vaatimuksista. Hieman huolestuttavaa oli varsinkin se, että yksi haastateltava piti Aluehallintoviraston tietämystä GDPR:n vaatimuksista puutteellisena. GDPR:n vaatimuksien tulkinnan haasteita esiteltiin tarkemmin luvussa 1.2.

> Terveydenhuollon organisaatiot ja niille töitä tekevät firmat tietävät aika hyvin, mutta sitten on harmaa-alue, jossa alkaa olemaan henkilökohtaisia tietoja, jotka ovat kriittisiä tai sensitiivisiä. Ne firmat ei vaan tiedosta sitä tai tiedostaa puutteellisesti, että heillä on tekeminen aika rempallaan, että isoillakin firmoilla on kummallisuuksia. (H2)

> Eihän hekään AVI:lla oikein tunnu ymmärtävän missä se data on ja kun miettii, että he on niitä ihmisiä, jotka antaa luvan käyttää palveluita tämmöisessä dataa hyödyntävässä kameravalvonnassa. (H3)

Kaikki haastateltavat olivat yhtä mieltä siitä, että olennaisimmat vaatimukset datalle liittyivät sen anonymisointiin, sijaintiin sekä tallennukseen ja säilyttämiseen dataprojektin alkuvaiheessa. Datan oli oltava anonymisoitua, jotta yksilöä ei pystytä siitä tunnistamaan mitenkään. Data näkyi sen käsittelijöille ainoastaan jonkinlaisena tunnuksena, mutta sen perusteella tietoja ei pystynyt yhdistämään tiettyyn henkilöön. Tietojen anonymisointia käytiin läpi tarkemmin luvussa 2.3.

> Datan oli oltava anonymisoitua, jotta emme varastoi nimiä tai muitakaan tunnistettavia piirteitä asiakkaasta. (H1)

> Eli tyypillisesti kaikki data mitä he jakavat meille täytyy olla anonymisoitu. (H4)

Kaikki haastateltavat kokivat, että datan sijainti oli sidosryhmien yksi keskeisimmistä vaatimuksista. Kokemukset siitä kuitenkin olivat hieman erilaisia. Yhden haastateltavan mielestä jotkin yritykset eivät ole vielä tarpeeksi valveutuneita tämän vaatimuksen suhteen, koska pilvipalvelun sijaintia, jossa data säilötään, ei kaikissa tapauksissa tiedetä. GDPR:n mukaan tietojen vapaan siirtämisen vuoksi datakeskuksien on sijaittava EU:n alueella (Euroopan parlamentti ja neuvosto 2016).

> Vaatimuksena oli, että meillä oli datakeskuksia maissa, joissa asiakkaita oli, joten data ei poistunut maasta. (H1)

> Eikä noista pilvistäkään aina tiedä, että onko ne edes EU:n alueella. Fiksut firmat ostavat pilvipalvelun niin et varmuudella tiedetään et konesali on tuossa. Mutta suurin osa ei ja se on just sitä et tekninen valveutuneisuus on niin heikkoa. (H2)

Asiakkaan vaatimus tietojen säilytyksestä määritellyn ajan puitteissa oli haastateltavien mielestä yksi olennaisimmista vaatimuksista. GDPR vaatimuksena on, että henkilölle, jonka tietoja käsitellään, on ilmoitettava hänen tietojensa säilytysaika (Euroopan parlamentti ja neuvosto 2016). Ei siis ole yllättävää, että asiakkaat vaativat tätä.

> Datan säilytysaika tuntui olevan yleinen vaatimus. Poistimme sen parin päivän kuluessa, emme oikeastaan tallentaneet dataa. (H1)

> Monelle isoin juttu tuntuu olevan se, että kuinka kauan niitä tietoja säilytetään, koska jonkun ajan päästä ne täytyy hävittää, jos ei niitä käytetä. (H3)

5.2 Tietojen läpinäkyvyys

Terveydenhuollon organisaatio on päävastuussa ilmoittamaan yksilöille heidän terveystietojensa käytöstä ja käsittelymenetelmistä. Tietojoukko oli myös palautettava asiakkaalle, jos he näin halusivat. Pk-yritykset ilmoittavat terveydenhuollon organisaatiolla kaikki tavat, joilla näitä tietoja jaetaan ja käytetään, jotta varmistetaan avoimuus ja väärinkäytösten estäminen. Yleisin tapa tehdä tämä oli luoda dokumentaatio, jossa selitettiin analytiikkaratkaisun datan käsittelyyn liittyvät asiat.

> Jotta saadaan suostumus et saadaan asentaa kamera tilaan, niin pitää olla kerrottuna et mitä se tekee, millä tavalla se tekee, mitä se näyttää ja mitä se ei näytä, ja mitä se kerää ja ei kerää. Suostumus tulee omaiselta tai siltä ihmiseltä, josta se data kerätään kameran kautta. (H5)

Haastateltavien kokemukset kyseisen dokumentaation merkityksestä olivat kuitenkin hieman erilaisia. Kun haastateltavilta kysyttiin dokumentointiin liittyvistä vaatimuksista, yksi haastateltava ei ollut kokenut mitään vaatimuksia, varsinkaan tekniseen dokumenttiin. Hän koki teknisen dokumentin olevan enemmänkin tietoa yrityksen sisälle, jotta muillakin olisi jatkossa tietoa ratkaisusta. Yksi haastateltava koki puolestaan melko yleiseksi sen, että yhteistyökumppanit haluavat nähdä yrityksen tiedonhallintamallin suunnitelman ja dokumentin, josta käy ilmi, että hallintamallin asiat ovat toteutuneet dataprojektin puitteissa.

> Ei oikeastaan ollut mitään vaatimuksia, mikä ei ole hyvä asia. Yleisesti dokumentoin sekä koodin että käytettävyysdokumentin vain tottumuksesta, mutta kukaan ei oikeastaan kertonut minulle sitä, että se on tehtävä. (H1)

> Moni yhteistyökumppani vaatii meiltä heille päin tiedonhallintamallin suunnitelmaa, eli melkeinpä sellaista laatukäsikirjaa mitä me seuraamme meidän tiedonhallintamallissamme. (H4)

Teknisen dokumentin sisällöstä haastateltavat olivat kuitenkin yhtä mieltä. Dokumentissa tuli kuvailla datan kaikki käsittely menetelmät, algoritmien tarkoitus sekä tulosmittarit. Asioita, joita esiteltiin luvuissa 3.4 ja 3.5 sisällytetään tekniseen dokumenttiin.

> Minulla oli tekninen raportti tai dokumentti, joka selitti algoritmin eri vaiheet, mitä se tekee ja miksi. Myös mitä varten ne ovat. Ja myös ohjeistusta miten käyttää kutakin eri vaihetta ja miten tulkita niitä. (H1)

> Tehdään dokumentointi eri vaiheista eli mistä data tuli, mitä sille tehtiin, miten se käsiteltiin, miten virheitä korjattiin, miten suodatettiin, millä menetelmällä analysoitiin, mitkä olivat tulosmittarit. Tämä tehdään kaikki tavallaan yhteen pötköön sitten aika usein, joka on käytännössä HTML tai PDF dokumentti. (H2)

> Kun prosessoidaan näitä heidän kuvia siis CT ja MRI kuvia niin tästä kerrotaan hyvinkin tarkasti sopimuksien yhteydessä olevissa dokumenteissa, eli että miten se anonymisoidaan, minkälaisia turvallisia yhteyksiä, koodausmenetelmiä, salausmenetelmiä käytetään siinä yhteyden luomisessa pilvipalvelimeen, missä meidän tekoälyalgoritmit pyörii ja miten ne otetaan takasin vastaan ja se että kuinka kauan me säilytetään asiakkaan prosessoituja datoja ja niin edelleen, tällaista läpinäkyvyyttä kyllä. (H4)

5.3 Tietojen laatu

Datan ymmärtämisvaiheesta haastateltavilla oli melko samanlaisia kokemuksia. Kun kysyttiin tietojoukkojen sopivuudesta ja käyttövalmiudesta, yleinen mielipide oli se, että dataa on aina muokattava, eikä se ole koskaan suoraan käyttökelpoista. Dataprojekteissa

eniten aikaa kuluu datan muokkaamiseen ja mallintamiseen (Berthold ym. 2010, 26). Data tarkastetaan ja mahdollisesti muokataan, jonka jälkeen asiakkailta kysytään tarkentavia kysymyksiä datapisteistä, kuten mahdollisista virhearvoista, mittaustekniikoista sekä siitä, mistä eri tietojoukoista data on yhdistelty, koska data on epäjohdonmukaista. Tästä muodostuu iteratiivinen prosessi, joka jatkuu datan laadun parannuksesta mallinnusvaiheeseen saakka. Dataa joudutaan jäsentelemään ja luokittelemaan paljon ennen kuin se on valmis mallinnettavaksi. Tätä voidaan kutsua myös datan annotoinniksi. Datan annotoinnin avulla pyritään myös karsimaan pois tietojen epäjohdonmukaisuutta.

> Kysyn asiakkaalta sen semantiikan sille datalle, että mitä arvoja ja virheitä näissä voi olla ja missä tilanteissa näitä tuli, sitten ajan sen datan läpi ja minulta tulee uusi kysymyssetti, koska sitten löytyy taas paljon lisää tapauksia, jotka ovat epäselviä, siitä tulee sellainen iteratiivinen prosessi. (H2)

> Omasta näkövinkkelistä lähtisin tarkastelemaan käytettävyyden kannalta, että saadaanko sitä tarpeeksi ja pystytäänkö siitä tekemään luotettavasti tietyt analyysit. (H3)

> Kun puhutaan näistä tekoälyalgoritmeista ja niiden kehityksestä niin ne ovat hyvin vaativia sen opetusdatan yhtenäisyyden suhteen, että datapisteet ovat oikein annotoitu. Ja sitä me joudumme tekemään aika paljon. (H4)

> Eihän se heti käyttökelpoista ole, kyllä siitä pitää tulkinta tehdä ensin. (H5)

Tietojoukot sisältävät melko poikkeuksetta aina epäjohdonmukaisia ja puolueellisia tietoja (Berthold ym. 2010, 39). Kun haastateltavilta kysyttiin tietojen epäjohdonmukaisuudesta sekä puolueellisuudesta, kaikkien kokemukset viittasivat siihen, että tietojoukot sisältävät aina tällaisia tietoja. Näitä tietoja on muokattava, jotta algoritmien antamat lopulliset tulokset olisivat mahdollisimman luotettavia. Datasta on tehtävä käyttökelpoisempaa poistamalla epäjohdonmukaisia, virheellisiä ja puolueellisia tietoja (Janssen ym. 2020). Haastateltavat kokivat ihmisen vaikutuksen tietojoukkoihin merkittävämpänä tekijänä puolueellisen tiedon muodostumiselle

> Aika usein data on epätasapainossa, siellä on eri määriä jotain tiettyä luokkaa siinä datassa ja käytännössä reaalimaailman datat ovat aina sellaisia. (H2)

> Ihmiset tekevät omista lähtökohdista erilaisia kirjauksia erilaisilla termeillä. Toiselle jalka poikki on sitä, että sillä on pikku haava. Toiselle se on sitä, että on täysin liikuntakyvytön, (H3)

> Ehdottomasti on puolueellista dataa, jos joku sanoo, että ei niin sitten ne ei ole ihmisten tekemiä tai vaikuttamia. (H4)

Tietojen ajankohtaisuuden koettiin olevan aina kunnossa. Usein haastateltavien projekteissa ainoastaan ajankohtaisella tiedolla pystyttiin tekemään sellainen ratkaisu, joka oli asiakkaalle hyödyllinen. Joissakin erikoistapauksissa voitiin hyödyntää myös vanhempaa tietoa, esimerkiksi kun pyrittiin toistamaan vanhaa hoitomuotoa. Ajankohtaista ja vertailukelpoista dataa on hankittava, jotta voidaan luoda relevantteja analytiikkaratkaisuja (Al-badi ym. 2018; Hripcsak ym. 2014; Janssen ym. 2020; Khatri & Brown 2010).

> Yhdessä projektissa analysoimme dataa reaaliajassa ja säilömme sitä pari päivää ja poistimme kokonaan, se oli todella uutta. Harvoin menimme taaksepäin katsomaan dataa tässä projektissa. (H1)

> Yleensä data on uutta. Mutta jos halutaan tehdä saman tyyppinen koe, kun joskus on tehty aiemmin niin pitää kaivaa sitä vanhaa dataa jostakin, ellei niitä revitä hatusta. Tämä on ihan fiksua, kun pitää kuitenkin pystyä arvioimaan hoidon tai menettelyn keskimääräinen vaikutus. (H2)

> Uudet anturit mitä on näissä hankkeissa hyödynnetty, niin se data mitä saadaan, on entistä tarkempaa, laaja-alaisempaa ja reaaliaikaisempaa. (H3)

> Kaikki kuvadata mitä kerätään niin se tyypillisesti alkaa olla vanhentunutta viimeistään 10 vuoden jälkeen. Viisi vuotta saattaa olla sellainen, että kuvat alkavat olla laadultaan huonompia kuin tänä päivänä kerätyt kuvat, koska teknologia kehittyy niin nopeasti. (H4)

Tietojen laadun parannukseen on olemassa useita eri keinoja. Esimerkiksi jonkun tietyn tekijän perusteella voidaan karsia joitakin tietoja pois. Tällaista voidaan tehdä, esimerkiksi luvussa 3.2 esitellyillä metodeilla. Yksi yleisimmistä keinosta korvata puuttuvia arvoja on korvata ne keskiarvolla, joka on laskettu kyseisen muuttujan datajoukosta. Näistä yleisimmistä käytetyistä keinoista haastateltavat olivat samaa mieltä.

> Laatua parannetaan yksinkertaisimmillaan niin, et yksittäisen attribuutin perusteella karsitaan pois sellaisia arvoja, jotka ylittävät tietyt minimi tai maksimi rajat. Jos puuttuu dataa, niin korvataan se keskiarvolla koko aineistossa, tämä on vanha kultanen standardi. (H2)

5.4 Tietojen laajuus

Haastateltavilta kysyttiin tässä osiossa tietojen laajuudesta ja sen vaikutuksesta dataprojektiin. Tietojen yhdistely eri lähteistä vaikuttaa merkittävästi datan laajuuteen. Haastateltavien kokemukset erosivat toisistaan jonkin verran, eikä selvää yleistä

näkemystä liittyen tietojen laajuuteen saatu muodostettua. Tähän myös varmasti vaikutti se, että haastateltavat ovat olleet mukana erilaisissa dataprojekteissa, joissa on kehitelty erilaisia ratkaisuja ja projekteilla on ollut erilaisia datatarpeita. Joissakin tapauksissa dataa oli yhdistelty useammasta eri tietolähteestä, joka sitten päätyi käyttöön.

> Joissakin tapauksissa asiakas on sitten tehnyt yhdistelemisessä sellaisia ratkaisuja mitkä eivät ole fiksuja. Jos yhdistelemisessä on vähänkin vapausasteita niin sitten kysytään, että kerrotteko, miten data on tarkalleen ottaen tuotettu ja millä kriteereillä yhdistetty. (H2)

Asiakas pyrkii siis useimmiten tekemään toimenpiteet datalle itse ja tämä saattaa aiheuttaa lisätyötä analyytikolle, koska data ei ole suoraan sopivaa projektin tarpeisiin. Toisissa tapauksissa käytettävälle datalle oli hyvin selkeä lähde, jolloin tietojen laajuuteen ei liittynyt erityistä kompleksisuutta.

> Olemme tarvinneet dataa vaan yhdestä tietolähteestä. Se on ollut suhteellisen yksinkertaista eksportoida yhdellä kertaa. (H4)

Kun haastateltavilta kysyttiin datan riittävästä määrästä, kokemukset olivat erilaisia. Haastateltavat kokivat, että dataa on ollut heidän projekteihinsa nähden tarpeeksi, liian vähän ja jopa ehkä hieman liikaakin.

> Kyllä sain aina dataa tarpeeksi enkä tarvinnut enempää. (H1)

> Valitettavan usein ei ole. Se on joko niin et datan tasapainoon liittyen meillä on liian vähän jonkun luokan edustajia siinä datassa, että pystyttäisiin tekemään siitä luotettavia päätelmiä tai sitten yksinkertaisesti dataa on vähän. (H2)

> Dataa alkaa olemaan niin paljon saatavilla, että sen oikea hyödyntäminen niin että se palvelisi, esimerkiksi hoitohenkilökuntaa, niin siinä meillä on aika paljon vielä tekemistä. (H3)

Haastateltavilla oli myös kokemuksia GDPR:n vaatimuksista liittyen tietojen laajuuteen. Jos prosessoidaan henkilökohtaisia arkaluontoisia tietoja, kuten terveystietoja, prosessointia varten voidaan käyttää ainoastaan sitä tietomäärää mikä on tarpeen. Ylimääräinen tieto on poistettava. (Euroopan parlamentti ja neuvosto 2016.)

> Me seuraamme tietosuojalain yhtä pykälää, jossa sanotaan, että meidän pitäisi ylläpitää kehityksessä vain sellaista määrää dataa, jota me tarvitaan ja kaikki muu ylimääräinen pitäisi poistaa. (H4)

> Ja tietoinen yhdistäminen se on taitaa olla se vaikein juttu siinä, mitä saa yhdistellä ja mitä ei. Et asiakaskin halusi käyttää sosiaalitunnusta, mutta se ei ollut pakollista eikä välttämätöntä meidän puolestamme. Mutta jos asiakas

> haluaa, niin voidaan käyttää. Tehdään kyllä selväksi, että se on sitten teidän vastuullanne, jos se näkyy jossain paikassa. (H5)

5.5 Tietovastaavan määrittäminen

Kun haastateltavilta kysyttiin tietovastaavan merkityksestä dataprojekteissa, vastaukset olivat melko erilaisia. Osalla ei ollut mitään tietoa kuka on ollut tietovastaavana dataprojekteissa. Osa puolestaan tiesi ja tiedossa olivat myös kriteerit tietovastaavan valinnalle. Jakautuneet kokemukset johtuvat luultavasti siitä, että haasteltavat ovat toimineet eri rooleissa. Haastateltavilla, jotka ovat enemmän tekemisissä pelkästään datan ja algoritmien kanssa, ei ollut mitään tietoa tietovastaavasta tai vain hyvin vähän. Esimerkiksi tietovastaavan nimi saattoi olla tiedossa, mutta vastaavan rooli tai vastuualueet olivat täysin pimennossa. Haastateltavilla, jotka ovat olleet enemmän dataprojektien hallinnoinnissa mukana, oli hyvinkin tietoa tietovastaavan vaatimuksista, roolista ja kompetenssista.

> Tämmöinen on ehkä vilahtanut jossakin dokumenteissa, että täytyy olla. Siinä on sitten joku henkilö, joka vaan nimetty siihen. (H2)

> Pelkästään firman vaatimukset, että on tietovastaavat ja muut paikallaan niin nekin ovat sellaisia asioita, että et onko kaikissa vieläkään välttämättä kunnossa? Ja sitten toinen asia on se, että vaikka on nimetty niin ymmärtääkö se ihminen mikä sen vastuu on? (H3)

> Tietovastaavan valinta perustuu ihan pätevyyteen, että täytyy olla tästä tietosuoja-alasta käytännössä vahva kokemus ja mielellään laki koulutus. (H4)

> Pitää olla perusymmärrys datan käytöstä, datan mahdollisuuksista ja jotain ymmärrystä laista myös ja GDPR:stä että se pystyy ja osaa. Jos joku soittaa ja haluaa tiedot poistettavaksi niin se tietovastaava sanoo, että ymmärrän. (H5)

Haastateltavien kokemuksista saattoi myös tehdä sen havainnon, että tietovastaavan määrittäminen koetaan pakollisena asiana. Eli joku henkilö on vain nimettävä tietovastaavaksi, koska regulaatio määrää niin, eikä tietovastaava välttämättä itsekkään tiedä mikä hänen vastuunsa on. Yrityksien on määriteltävä tietovastaava, jotta tietojen hallinnan vastuuvelvollisuus on virallista (Euroopan parlamentti ja neuvosto 2016). Tietovastaavan on sitouduttava ja ymmärrettävä hänen velvoitteensa liittyen terveystietoihin (Hripcsak ym. 2014).

5.6 Tietojen prosessointi

Haastateltavat eivät kokeneet erityisiä vaatimuksia datan prosessointimetodeihin liittyen. Asiakkaat olivat tyytyväisiä ratkaisuun, joka toimii heidän tarpeisiinsa.

> Asiakas haluaa vain ratkaisun toimivan. Käytimme vain yleistä koneoppimismallia tarkoituksiimme. Se on malli, jota ei ole erityisesti suunniteltu meidän käyttöön, mutta on samassa spektrissä. (H1)

> Aika usein siihen härveliin mikä laskee tai tekee jotakin niin siihen asiakas ei kyllä juuri mitään ole sanonut ikinä. (H2)

Sen sijaan lopullisten ratkaisujen kehittäminen ja luominen niin, että se on täysin GDPR regulaatioiden mukainen koettiin hyvinkin ongelmalliseksi. Kaksi haastateltavista ilmaisivat voimakkaasti oman mielipiteensä GDPR:n vaatimukseen, jonka mukaan yksilöiden tiedot on poistettava välittömästi tai mahdollisimman pian sopimuksen puitteissa, jos yksilö haluaa, että hänen tietojaan ei enää prosessoida. Oikeus tulla unohdetuksi tarkoittaa, että rekisterinpitäjän on poistettava yksilön, eli rekisteröidyn tiedot ilman turhaa viivytystä. (Euroopan parlamentti ja neuvosto 2016.) Tämän vaatimuksen täyttäminen koettiin erityisen ongelmallisena ja yrityksen toimintaa kuormittavana, jos tällaiseen tilanteeseen joudutaan.

> Mitä jos se henkilö sanoo et ei saakaan enää käyttää? Tämän leipominen kaikkeen niin et se oikeasti, ensinnäkin teknisesti hävitetään se tietue sieltä jostakin ja sitten kaikki tuotokset mitä on koskaan laskettu hävitetään, sen kontribuutio niihin kaikkiin on kyllä mielenkiintoinen. Esimerkiksi että joudutaan laskemaan jotkut helvetin isot mallit aina uusiksi sen takia. Tämän pitäisi olla niin kuin mahdollista. (H2)

> Jos sitä kerätään muutenkin sitä tietoa, niin millä helvetillä tehdään sen niin et pystytään poistamaan tiettyä henkilöä koskevat tiedot kameratallenteessa? Ei siellä lue, että on henkilön x tiedot. Siellä on kamera tallenne, jossa juoksen edes takasin, millä poistat ne jutut mitkä koskevat minua? Tuossa tulee vähän sellaisia tiettyjä hankaluuksia, että vedetäänkö tuossa organisaation narua liian tiukkaan kaulaan, että millä pystytään toteuttamaan tuommoisia vaatimuksia, on aika haasteellisia tehtäväksi. (H3)

Haasteltavien kokemukset analytiikkaratkaisujen validoinnista olivat pitkälti samanlaisia. Esimerkiksi mallien vertailua, mallin tarkkuuden ja mallin tarkkuuden epävarmuuden laskentaa tehtiin poikkeuksetta aina. Yksi keino on myös luoda verrokkidata (englanniksi *ground truth*), joka koostuu todellisista arvoista, jotka kuvaavat haluttua ilmiötä. Tätä dataa ei tietenkään opeteta kehitetylle mallille, koska se johtaisi muuten ylisovitukseen. Mallin ylisovituksella tarkoitetaan sitä, että malli ennustaa liian tarkasti tiettyjä

tietojoukkoja, eikä sovellu muiden tietojoukkojen ennustamiseen (Berthold ym. 2010, 102). Lopullista testataan tähän verrokki dataan, jotta saadaan mallin tarkkuus todellista dataa vasten. Se, onko tarkkuus tarpeeksi asiakkaan mielestä heidän tarpeisiinsa, on asiakkaasta itsestään kiinni. Asiakkaalle ilmoitetaan dataprojektin dokumentoinnissa mallin tarkkuudesta ja arvo tarkkuuden epävarmuudelle. Mallin hyväksymiskriteereistä viestitään myös asiakkaille ja sääntelijöille, jotta sidosryhmät voivat arvioida, onko mallin suorituskyky tarpeeksi hyvä kliiniseen käyttöön.

> Testasin sitä ja koska minulla oli entuudestaan kokemusta koneoppimisesta, vertasin siitä muihin malleihin. Pitäisi kuitenkin tehdä enemmän testejä. (H1)

> Aika usein lasketaan ratkaisun suorituskyky, pystytään laskemaan jotkin marginaalit simuloimalla tai dataa veivaamalla, jolloin sitten saadaan hyvinkin usein jokin tarkkuus ja sitten meillä on jokin epävarmuus sille tarkkuudelle. Voidaan tehdä epävarmuusajoja, että missä rajoissa joku vastaus on eli vastaus voi olla X mutta se voi olla myös + - Y jotakin 95 prosentin todennäköisyydellä. (H2)

> Tehdään niin, että meillä on ensin kehitysversio jostakin mallista ja oikeastaan jo siinä vaiheessa me on luotu verrokki data setti mikä on englanniksi ground truth, joka on se osa datasta mitä ei sille algoritmille näytetä. Algoritmia sitten ajetaan sitä vasten ja katsotaan että minkälaisia tuloksia se sai siihen ground truth dataan ja sitten määritellään sellaiset hyväksyntäkriteerit sekä itsellemme että ulospäin sääntelijöille, mitkä pitää läpäistä, että tämä mallin testaus hyväksytään ja että sen suorituskyky on tarpeeksi hyvä kliiniseen käyttöön. (H4)

Haastateltavat kokivat, että datan prosessointimetodien reiluus varmistetaan jo oikeastaan datan valmisteluvaiheessa, kun poistetaan virheellisiä, epäjohdonmukaisia ja puolueellisia tietoja. Kahdella haastateltavalla oli kuitenkin mielessä mahdollinen skenaario, jossa prosessoinnin reiluus ja oikeudenmukaisuus ei välttämättä toteudu, vaikka tietojen käsittelyssä ja laadussa olisi huomioitu GDPR:n vaatimukset sekä virheellisten ja puolueellisten tietojen poisto. Data on hyvin tunnistettavaa monella tavalla ja on mahdollista, esimerkiksi kuvadataa ehostamalla ja oikeita työkaluja käyttämällä luoda siitä tunnistettavampaa, jolloin pystytään tunnistamaan jopa yksilöitä.

> Vois ajatella et ei nyt nämä jonkun aivokuvat hirveästi kerro mitään, mutta on löytynyt jo tämmöisiä konsteja että, otetaan aivokuvat, tehdään kasvojen rekonstruktio niiden aivokuvien perusteella ja syötetään se kasvokuva kasvojentunnistus järjestelmälle ja pystytään tunnistamaan et kenen kuvat ne oli. On aika monimutkaisia nämä ketjut missä saat tämmöisiä pieniä vihjeitä, joilla pystyt jollakin tavalla ehostamaan sitä dataa ja mahdollisesti yhdistelemään niitä ja päättelemään kaikenlaista. Nämä on aika hurjia. (H2)

> Käyttäen sitä kehitettyä algoritmia niin pystytäänkö me sitä tutkimalla jonkun datajoukon kanssa jotenkin tunnistamaan jokin kuva et mitä on käytetty sen algoritmin opetukseen ja sitä kautta saamaan tietoa, että tämmöinen henkilö on ollut tämän algoritmin kehitysjoukossa. (H4)

5.7 Tiedonhallintamallin vaikutus dataprojekteihin

Haastateltavilta kysyttiin mielipiteitä kirjallisuuskatsauksen perusteella luotuun tiedonhallintamallin viitekehykseen sekä yleisesti tiedonhallintamallista. Haastateltavat olivat melko yksimielisiä siitä, että kaikki tiedonhallintamallin osat ovat hyödyllisiä ja ne antavat hyvät lähtökohdat dataprojektin läpivientiin. Tiedonhallintamalli koettiin erityisen merkittävänä dataprojekteille, koska se hidastaa kehitystyötä. Tiedonhallintamallia tulisi jatkuvasti pyrkiä parantamaan, jotta kehitystyö sujuisi paremmin.

> Tämä avaa aika paljon näitä erinäköisiä asioita mitä pitää ottaa huomioon tai mitä olisi hyvä ottaa huomioon, kun aloittelee tekemään jotakin. Nämä on nimenomaan kyllä erittäin hyviä. (H2)

> Joo tosiaan tämä teema tiedonhallintamalli, niin se on erittäin tärkeä ja itseasiassa se on yks sellainen asia mikä on selkeästi meidänkin firmassa tällä hetkellä pullonkaula meidän kehitystyössä, eli meillä on huomennakin palaveri siitä että miten meidän pitäisi parantaa tätä datan hallintaa ja mallia niin että me itse tiedetään helposti mitä datoja meillä on ja sitä on helppo etsiä ja että tiedetään niiden lähteet ja niin edelleen. (H4)

Haastateltavat kokivat myös tiedonhallintamallin hyödyn pk-yritykselle merkittävänä. Erityisesti silloin kun yritys on kasvamassa ja hallittavia tietojoukkoja alkaa olemaan huomattavan suuri määrä, tämän viitekehyksen asioita tulisi laittaa kuntoon, jotta asiakkaiden vaatimuksiin voidaan vastata ja jotta dataprojektien läpivienti onnistuisi paremmin. Tiedonhallintamalli koettiin myös sellaisena asiana, joka pitäisi sisällyttää yrityksen ydintoimintaan, jotta asiat tehdään oikealla tavalla. Tiedonhallintamallin implementoinnin hyötyjä pk-yrityksille avattiin tarkemmin luvussa 2.3.

> Nämä on kaikki relevantteja asioita ottaen huomioon organisaation, mutta varsinkin organisaation joka kasvaa eikä startupille, joka on pienimuotoinen koska siinä vaiheessa ei ole aikaa ajatella näitä. Siinä vaiheessa kun firmalla on yhteistyökumppaneita kymmenkunta ja enemmän, datoja on tuhansia tai sanotaan että datapisteitä on kymmeniä tuhansia ja työntekijöitä on yli kymmenen niin siinä vaiheessa kyllä näitä asioita pitää alkaa miettimään ja pistämään paikalleen. Big data on pyörinyt kaikkien huulilla viimeset 10 vuotta ja ollaan vasta käytetty se jäävuoren huippu siitä, jotta me saadaan koko vuori käytettyä niin tarvitaan aika paljon tehokkaammat ja paremmat prosessit ja menetelmät tällaiselle tiedonhallintamallille. (H4)

> Siis tämän periaatteessa pitäisi olla yrityksen DNA:han rakennettu. Jos ruvetaan käsittelemään henkilöihin liittyvää dataa niin pitää tehdä asiat by the book. Ehdottomasti siinä mielessä pitäisi olla tämmöinen ja pitäisi olla kirjattuna projektiin kun tehdään, koska jos joku kysyy et miten nämä asiat on otettu huomioon niin voidaan sanoa, että ihan suunnittelusta asti on huomioitu. (H5)

Vaikka tiedonhallintamallin viitekehyksen osat koettiin relevantiksi ja vaatimustenmukaisuuteen liittyen tärkeäksi, mallin rakenne koettiin myös raskaaksi toteuttaa, mikä johtui pitkälti GDPR:n vaatimusten toteuttamiseen, joihin monet tiedonhallintamallin osat toki viittaavat. Varsinkin GDPR:n vaatimukset oikeudesta tulla unohdetuksi sekä tietojen läpinäkyvyys liittyen prosessointimetodeihin koettiin ongelmallisina toteuttaa. Prosessointimetodien läpinäkyvyys ei välttämättä ole hyödyksi yksilölle, mutta sidosryhmiin kuuluvat asiantuntijat ja tarkastajat voisivat tarkistaa algoritmien toiminnan puolueellisia päätöksiä tekevän analytiikkaratkaisun varalta (Janssen ym. 2020). Tämä herätti varsinkin huolta yhden haastateltavan osalta, joka on toiminut dataprojektien hallinnoinnissa tietoturva-asioissa.

> Onhan tuossa paljon ja ehkä liikaakin, sen takia olen tietyllä tapaa noista varovainen et ei tehdä omasta elämästä liian vaikeata. Jos jakaa jotain tietoa, niin käy lähinnä rekisteriselosteet, vastuuhenkilöt, tämän tyyppiset asiat et näiltä henkilöiltä saa vastauksen, jos on kysyttävää. Nämä on meidän politiikat näissä asiakkaan suuntaan, että ei avattaisi ihan mitään käytännön ratkaisuja asiakkaille, ainakaan kooditasolle. Jos miettii vakoilua tai haitantekoa, ettei niitä liikaa kerro asiakkaalle. (H3)

Sen lisäksi, että tiedonhallintamallin katsottiin olevan sekä hyödyllinen, että jossakin määrin vaikeasti toteutettava pk-yritykselle, myös sidosryhmien tiedonhallintamalleihin toivottiin parannusta. Erityisesti terveydenhuollon organisaatioiden tiedonhallintamallien kypsyys herätti keskustelua haastateltavien kanssa, koska hyvä asiakkaan hallintamalli hyödyttää myös yritystä. Tällöin, esimerkiksi tietojen laatu sekä tietoturvallisuus paranisivat ja yleisesti kommunikointi olisi helpompaa. Terveydenhuollon organisaatioiden tiedonhallintamalli vaatii vielä joiltakin osin paljon työtä, eikä sen toteutus ole vielä toistaiseksi, varsinkaan tietojen laadun suhteen ole onnistunut.

> Tulee vähän sellainen fiilis et joku on sanonut heille, että sano näin. Monesti nämä vaatimukset on sen tasoisia, että jotenkin aistii läpi et ne ei itsekkään ymmärrä mistä on kyse välttämättä, se on kyllä aika usein sellainen mihin on törmännyt. (H3)

> Jos ne nyt sen osaa et pidetään omista salasanoista huolta, niin se on aika paljon. Varsinkaan kentällä tietojen hallinta ei ole korkealla tasolla. (H3)

Heidän pitäisi huolehtia siitä, että nämä datat on järjestetty hyvin, niitten metadata olisi merkattu ja tallennettu fiksusti niin että tätä dataa vois käyttää hyväksi tulevaisuudessa hyödyllisten algoritmien kehittämisessä. Mutta luulen että tiedonhallintamallin aikaansaamiseksi ihmiset joutuvat käyttämään hirveästi aikaa metadatan luomiseen, ja tämä taitaa olla se pullonkaula. (H4)

Vaikka lääkärit on jo halunnut tehdä tällaisen tietoaltaan ja hyvän tiedonhallintamalli jutun niin ne lääkärit on joutuneet merkkaamaan siinä kliinisessä työssä näitä datoja aika paljon suuremmalla mittakaavalla, kun aiemmin, jolloin he sanoo tekevänsä kolmasosa enempi töitä sen takia, jotta tulevaisuudessa voitaisiin kehittää näitä algoritmeja ja mikä on ajanut siihen, että ne lääkärit on lähtenyt sieltä yliopistosairaalasta pois, kirjaimellisesti ottanut loparit. (H4)

6 Tiedonhallintamalli terveysteknologia-alan pk-yrityksissä

6.1 Johtopäätökset

Haastateltavien kokemukset tiedonhallintamallien osista ja niiden sisällöstä vastasivat pitkälti teoreettisessa viitekehyksessä määriteltyjä tiedonhallintamallin osia. Haastateltavat kokivat viitekehyksen osa-alueet tarpeellisiksi ja tärkeiksi asioiksi, joita erityisesti kasvavan pk-yrityksien on laitettava kuntoon terveysdatan käsittelyyn liittyen, jotta voidaan vastata asiakkaiden vaatimuksiin ja dataprojektien läpivienti onnistuisi paremmin. Tiedonhallintamallin puutteen koettiin muodostavan erityisen hidasteen dataratkaisujen kehitystyölle, jonka vuoksi tiedonhallintamallin käyttöönotto koettiin erityisen tärkeänä. Mielipiteitä ilmeni myös siitä, että viitekehys on mahdollisesti liian työläs toteutettavaksi, varsinkin pk-yrityksen rajalliset resurssit huomioiden. Tiedonhallintaan liittyvät vaatimukset ovat usein merkittäviä, jolloin tiedonhallintamallin toteuttaminen ei ole mieluisaa yrityksille (Begg & Caira 2011). Yritykset kokevat regulaatioiden noudattamisen liian kuormittavana (Sirur ym. 2018). Toisaalta muodostetun tiedonhallintamallin osat perustuvat pitkälti GDPR:n vaatimusten noudattamiseen, joten näitä asioita tulee noudattaa joka tapauksessa huolimatta siitä, kokeeko pk-yritys sen työläänä vai ei. Haastatteluissa nousikin esiin kommentteja, joissa toivottiin päättäviltä tahoilta järjenkäyttöä liittyen vaatimusten laatimiseen, jotta yrityksien työ vaatimustenmukaisuuden hyväksi ei ole kohtuuton. Haastateltavien kokemuksia tiedonhallintamallien vaikutuksesta dataprojekteihin terveysteknologia-alalla käytiin läpi luvussa 5.7.

GDPR:n vaatimusten noudattaminen koettiin raskaana osittain siitä syystä, että vaatimukset ovat liian haastavia. Erityisen mielenkiintoisia mielipiteitä herättivät GDPR:n vaatimukset yksilön oikeudesta tulla unohdetuksi, prosessointimetodien reiluudesta ja tietovastaavan määrittämisestä. Yksilön oikeus tulla unohdetuksi koettiin erityisen vaikeana toteuttaa, koska ratkaisuja kehitettäessä tämä ei ole juurikaan keskiössä. Haastatteluiden perusteella yksilöiden tietoa koskevat vaatimukset painottuvat pitkälti datan anonymisoinnin ympärille. Tilanne, jossa yksilö voisi vaatia tietojensa poistamista kaikista mahdollisista käsittelyprosesseista koettiin kestämättömänä yritykselle, koska spesifisti juuri tiettyä henkilöä koskevien tietojen jäljittäminen malleissa ja palvelimissa veisi tolkuttoman paljon aikaa. Tähän olisi vielä lisättävä isojen datamallien laskeminen uudestaan, kun yksilön tiedot on poistettu. Haastateltavien

terveysdataprojekteissa ei kuitenkaan ole ollut toistaiseksi tarvetta ryhtyä tällaisiin toimenpiteisiin. Terveysteknologiayritysten olisi kuitenkin syytä varautua tällaiseen tilanteen varalta, esimerkiksi huomioimalla tämä asia paremmin analytiikkaratkaisun suunnittelussa ja tietojoukkojen luomisessa.

Haastatteluiden perusteella prosessointimetodien reiluuden katsottiin olevan kunnossa, jos datan virheellisyyteen ja puolueellisuuteen on puututtu riittävästi datan valmisteluvaiheessa. Kun datasta on poistettu nämä tiedot, analytiikkaratkaisun tulokseen voidaan luottaa paremmin (Janssen ym. 2020). Ratkaisujen tulisi olla reiluja ja oikeudenmukaisia yksilön näkökulmasta (Euroopan parlamentti ja neuvosto 2016). Osa haastateltavista piti kuitenkin mahdollisena sitä, että vaikka yksilön tunnistaminen datasta olisi minimissä, oikeilla työkaluilla olisi mahdollista silti selvittää yksilön tiedot. Tämä näkemys tosin koski vain yksilöistä kerättyä kuvadataa, esimerkiksi MRI kuvia. Teoriassa tämä olisi kuitenkin mahdollista, esimerkiksi jos mallin opetukseen käytettyä kuvadataa päätyisi vääriin käsiin ja tätä kuvadataa ehostettaisiin ja syötettäisiin kasvojen rekonstruktio työkaluihin. Lopputuloksena saattaisi olla kuva, josta yksilön voi tunnistaa. Tämän ehkäisemiseksi tiedonhallintamallin on oltava kunnossa, jotta dataprojekteissa käytettävät tietojoukot on turvallisesti säilötty sekä vastuut niiden hallinnasta ovat selvillä, eikä luvattomilla henkilöillä ole pääsyä niihin.

Tietovastaavan määrittäminen todettiin teoriaosuudessa yhdeksi merkittäväksi tiedonhallintamallin osaksi. Tietojen hallinnan vastuuvelvollisuus on oltava virallista, joten yrityksien on määriteltävä tietovastaava (Euroopan parlamentti ja neuvosto 2016). Terveystietojen hallintaan liittyy merkittävä vastuu, joten tietovastaavan on sitouduttava ja ymmärrettävä hänen velvoitteensa (Hripcsak ym. 2014). Haastateltavien kokemuksista pystyi kuitenkin tekemään havainnon, että tietovastaavaksi vain valitaan joku henkilö, koska niin on tehtävä. Haastatteluissa ilmeni myös näkemyksiä siitä, että tietovastaava ei välttämättä itsekään ymmärrä omaa vastuutaan terveystietojen hallintaan liittyen. Tietovastaavan kompetenssiin varmistamiseen ja perehdytykseen tulisi panostaa terveysteknologiayrityksissä enemmän. Haastatteluissa ilmenneitä asioita tietovastaavan määrittämiseen liittyen käytiin läpi luvussa 5.5.

Haastateltavien kokemukset sidosryhmien tiedonhallintamallin puutteesta ja jonkin asteinen tietämättömyys GDPR:n vaatimuksista herättivät jopa suoranaista huolta. Haastatteluissa kävi ilmi, että pk-yrityksillä on haasteita toimia vaatimusten mukaan

oikein, koska päättävät tahotkaan eivät välttämättä ymmärrä vaatimustenmukaisuutta syvällisesti. Tämä saattaa johtaa, esimerkiksi merkittäviin tietosuojariskeihin, jotka koskevat yksilöiden terveystietoja ja myöhemmässä vaiheessa riskeihin puuttuminen on vaikeampaa. Tämä on hyvin ongelmallista pk-yrityksen kannalta, koska oletettavasti varsinkin päättävien tahojen on tarkalleen tiedettävä regulaatioiden vaatimukset.

Yksi haastateltava puolestaan koki terveydenhuollon organisaation tiedonhallintamallin puutteen vaikuttavan suoraan yrityksen dataprojektien toteutumiseen, mikä tietysti pitää paikkaansa. Jos asiakkaalta ei saada tarpeeksi laadukasta, järkevästi muodostettua, tarkkaa ja puolueetonta dataa, hyödyllisen data-analytiikkaratkaisujen luominen muuttuu todella haastavaksi. Tämä voi johtaa siihen, että kehitetyt data-analytiikkaratkaisut voivat tehdä päätöksiä, jotka eivät ole yksilön edun mukaisia (Winter 2021). Toisaalta ihmisten vaikuttamien tietojoukkojen koettiin aina sisältävän virheellistä ja puolueellista tietoa. Haastateltavien mukaan reaalimaailman datat ovat aina tällaisia. Datajoukot sisältävät lähes aina epäjohdonmukaista ja puolueellista dataa (Berthold ym. 2010, 39). Jos tiedonhallintamalli olisi paremmin kunnossa terveydenhuollon organisaatiossa, tietojen laatu olisi parempi ja yrityksen ei tarvitsisi käyttää niin paljon aikaa datajoukkojen valmisteluun. Terveystietojen hyödyntäminen terveysteknologiassa vaatii siis myös terveydenhuollon organisaation panostusta tiedonhallintamalliin, jotta tietojoukkojen laatu pysyy korkeana koko ketjun ajan, sen keräämisestä hyödyntämiseen. Haastatteluissa ilmenneitä tietojen laatuun liittyviä asioita käytiin läpi luvussa 5.3.

Haastatteluiden avulla testattu tiedonhallintamallin viitekehys oli siis suurimmalta osin hyvä. Osa-alueita ei haastatteluiden perusteella tarvitse muuttaa. Sen sijaan, joitakin osa-alueiden määritelmiä olisi syytä tarkentaa hieman. Esimerkiksi haastatteluissa kävi ilmi, että tietojen säilytysaika on merkittävä vaatimus, mistä terveydenhuollon organisaatio vaatii selvitystä. Tämä olisi syytä nostaa enemmän esiin viitekehyksessä. Myös tietovastaavan perehdytystä olisi syytä korostaa viitekehyksessä enemmän, koska haastatteluiden perusteella tietovastaava ei välttämättä ymmärrä vastuutaan ja hänet nimetään vain koska on pakko. Alla olevassa taulukossa (Taulukko 9.) on näiden lisäksi tarkennettu myös lisää sidosryhmien vaatimuksia, joita käytiin läpi luvussa 5.1, tietojen läpinäkyvyyttä, jota käytiin luvussa 5.2 sekä tietojen prosessointia, jota käytiin läpi luvussa 5.6.

Taulukko 9. Tarkennettu tiedonhallintamallin viitekehys

Osa	**Määritelmät**	**CRISP-DM-vaihe**
Sidosryhmien vaatimukset	Yrityksien ja terveydenhuollon organisaatioiden oikeudet ja velvollisuudet terveystietojen keräämisen ja käytön kannalta on ymmärrettävä GDPR:n ja tietosuojan suhteen. **Datan täytyy olla anonymisoitua.** **Palvelimien on sijaittava EU:n alueella.** **Dataa on säilytettävä vain sen aikaa, kun on tarpeen.**	Liiketoiminnan ymmärtäminen, arviointi, käyttöönotto
Tietojen läpinäkyvyys	Ihmisille ja organisaatioille tulee ilmoittaa, jos heidän tietojaan jaetaan ja käytetään avoimuuden varmistamiseksi ja väärinkäytösten estämiseksi. **Yrityksissä on oltava vakiintunut tapa dokumentoida tietojen käsittelyyn ja mallinukseen liittyvät vaiheet, jotta tiedot ovat läpinäkyviä sisäisesti ja tarvittaessa ulkoisesti.**	Liiketoiminnan ymmärtäminen, arviointi
Tietojen laatu	Dataa on saatettava käyttökelpoiseen muotoon poistamalla epäjohdonmukaisia, virheellisiä ja puolueellisia tietoja. Datan oltava vertailukelpoista, ajankohtaista ja saatavilla, jotta voidaan tuottaa luotettavia analytiikka ja tekoälyratkaisuja. Datan laatua on seurattava koko projektin ajan.	Datan ymmärtäminen, datan valmistelu, datan mallinnus, arviointi
Tietojen laajuus	Eri tietolähteiden selvitys ja yhdistäminen, jotta tietojoukot eivät ole puutteellisia ja varmistetaan käytettävien prosessointi tapojen riittävyys. Jaetaan vain tarpeellinen tieto.	Liiketoiminnan ymmärtäminen

Osa	Määritelmät	CRISP-DM-vaihe
Tietovastaavan määrittäminen	Yrityksien on määriteltävä tietosuojavastaava GDPR:n vaatimusten mukaan, jotta tietojen hallinnan vastuuvelvollisuus on virallista. **On varmistettava, että tietovastaava ymmärtää vastuunsa ja että hänen roolinsa on näkyvä yrityksessä.**	Liiketoiminnan ymmärtäminen
Tietojen prosessointi	Datan prosessointimetodit on määriteltävä niin, että ne ovat reiluja ja asianmukaisia yksilön suhteen. **Prosessointimetodien suorituskyky ja epävarmuus on laskettava.** **Suorituskyvyn on ylitettävä määritellyt rajat, jotta malli sopii kliiniseen käyttöön.**	Datan mallinnus

6.2 Yhteenveto

Henkilökohtaisten terveystietojen hyödyntäminen data-analytiikassa on kasvanut viime vuosina huomattavasti, koska terveystietojen määrä on kasvanut erilaisten lähteiden myötä. Näitä lähteitä ovat, esimerkiksi monipuolisemmat tietojärjestelmät, sensorit ja päätelaitteet. Kasvavat tietomäärät useista lähteistä aiheuttavat terveystietojen kokoamisvaiheessa tietojen epäjohdonmukaisuuksia. Tiedonhallintamallin tarkoituksena on varmistaa yrityksen toimintatavat tietojen käsittelyn suhteen, jotta lisätään datasta hyödynnettävää arvoa ja vähennetään tietojen hallintaan liittyviä kustannuksia ja riskejä. Regulaatiotahot painottavat myös tiedonhallintamallin olemassaoloa yrityksissä nykyään enemmän, jotta terveystietojen käsittelyyn ei liity väärinkäytöksiä. Kehittämällä tiedonhallintamallia, yritykset todistavat olevansa päteviä käsittelemään terveysdataa systemaattisesti regulaatioiden ja säädösten mukaan. Tiedonhallintamalleja ja terveysteknologia-alalle soveltuvia viitekehyksiä on tutkittu aiemmin hyvinkin runsaasti, mutta varsinkaan pk-yrityksille sopivia tiedonhallintamalleja terveysteknologia-alalle ei ole juurikaan tutkittu. Oletus aikaisemmassa tutkimuksessa on ollut se, että minkä tahansa tiedonhallintamallin pystyy skaalaamaan pk-yritykselle sopivaksi käyttöön.

Tutkimuksessa muodostettiin kolme tutkimuskysymystä, joiden kautta tiedonhallintamallin vaikutusta terveysteknologia-alan pk-yrityksiin lähdettiin tarkastelemaan. Ensimmäisessä tutkimuskysymyksessä selvitettiin, *Mitä hyötyä tiedonhallintamallin toteuttamisesta on terveysteknologia-alalla toimiville pk-yrityksille.* Tähän vastausta on käsitelty luvussa 2.3, jossa esille nousi erityisesti vaatimustenmukaisuus ja miten vaatimustenmukaisuus voidaan saavuttaa sisäänrakennetulla tiedonhallintamallilla niin, että pk-yritys pystyy jatkossakin toimimaan vaatimustenmukaisena. Haastatteluissa nousi samoja piirteitä esiin kuin teoriassakin. Erityisesti tiedonhallintamallin viitekehyksen asioiden tärkeyttä ja niiden kuntoon laittamista korostettiin pk-yrityksiltä, koska muuten toiminta terveysteknologia-alalla ei ole mahdollista ongelmitta tai ylipäätänsä ollenkaan. Haastatteluissa kävi myös ilmi, että asiakkaat voivat vaatia yrityksiltä tietojen käsittelyyn liittyvää dokumentaatiota, esimerkiksi tiedonhallinnan käsikirjaa, jossa kerrotaan mitä asioita yritys huomio tiedonhallintamallissa. Tämä korostaa myös tiedonhallintamallin tarpeellisuutta pk-yrityksissä.

Toisessa tutkimuskysymyksessä tutkittiin, *mitä osa-alueita tiedonhallintamallin pitää sisältää, kun pk-yritykset valmistautuvat käyttämään data-analytiikkaa terveystietojen käsittelyä varten.* Terveysdataan ja terveysteknologia-alalle sopivia tiedonhallintamallin osa-alueita tarkastellaan pitkälti teoriaosuudessa luvussa 2. Osa-alueiden määrittelyssä pyrittiin huomioimaan pk-yrityksen tarpeet, jotta tiedonhallintamallin ehdotetusta viitekehyksestä ei tule liian raskasta. Ehdotettu tiedonhallintamallin viitekehys tiedonhallintaan terveysteknologiayrityksille (ks. Taulukko 6.) muodostettiin kirjallisuuskatsauksen pohjalta luvussa 2.5. Tätä viitekehystä testattiin empiriaosiossa. Testin tuloksena tiedonhallintamallin joitakin osia tarkennettiin määrittelyjen osalta. Tarkennettu tiedonhallintamallin viitekehys on esitetty luvussa 6.1 (ks. Taulukko 9.).

Kolmannessa tutkimuskysymyksessä selvitettiin *miten tiedonhallintamalli vaikuttaa terveysteknologia-alan data-analytiikkaprojektien toteutukseen*. Tiedonhallintamallin vaikutusta data-analytiikkaprojektien CRISP-DM-prosessiin selvitettiin luvussa 3. Tiedonhallintamallin viitekehyksen osien katsottiin vaikuttavan monipuolisesti jokaisen CRISP-DM-prosessivaiheen onnistumiseen. Tiedonhallintamallien osien vaikutus eri prosessivaiheisiin kerättiin taulukkoon (ks. Taulukko 7.) luvussa 3.7. Teemahaastattelun rakenne ja haastattelukysymykset perustuvat myös tähän taulukkoon. Haastattelun tulokset on kerätty kappaleeseen 5. tiedonhallintamallin osien mukaan. Haastateltavien

kokemusten perusteella nämä osa-alueet vaikuttavat aina dataprojekteissa, eivätkä dataprojektit onnistu ilman osa-alueiden noudattamista. Hyvä tiedonhallintamalli luo pohjan analytiikan käyttöönottoon, jolloin myös analytiikkaratkaisun tuottamiin tuloksiin voidaan luottaa paremmin (Okoro 2021). Haastatteluissa kävi myös ilmi, että yritykset eivät huomioi näitä kaikkia osa-alueita tasapuolisesti, mikä on toki ymmärrettävää, koska kaikki osa-alueet eivät vaikuta suoranaisesti dataratkaisun laatuun, kehitystyöhön ja valmistumiseen. Kaikki osa-alueet on kuitenkin toteutettava vaatimusten vuoksi ja yritykset myös tekevät niin.

6.3 Tutkimuksen luotettavuuden arviointi

Kvalitatiivissa tutkimuksissa tutkijan puolueettomuutta tulisi arvioida huolella, koska kvalitatiiviset tutkimukset perustuvat osittain tutkijan subjektiiviseen näkemykseen (Tuomi & Sarajärvi 2018). Tämän tutkimuksen ei voida katsoa olevan täysin puolueeton, koska tutkija työskentelee terveysteknologiayrityksessä, jossa myös yksi haastateltava tuotejohtaja työskentelee. Tutkimusta ei kuitenkaan ole tehty toimeksiantona tälle yritykselle, vaan yritys on ollut mukana antamassa kontribuutiota tutkimuksen aiheeseen, joka toki kiinnostaa yritystä. Samanlaisen kontribuution tutkimuksen aiheeseen ovat antaneet myös muut haastateltavat. On kuitenkin huomioitava, että tutkijan yhteydellä terveysteknologiayritykseen on varmasti ollut jotakin vaikutusta tutkijan subjektiiviseen näkemykseen. Tutkija on tästä yhteydestä huolimatta pyrkinyt olemaan mahdollisimman objektiivinen tutkimuksen toteutuksessa sekä aineiston keräämisessä ja analysoinnissa.

Tutkimuksen luotettavuutta lisää se, että haastateltavia on pyritty valitsemaan niin kehityspuolelta, kuin myös projektien hallinnointipuolelta. Tuloksia saatiin tämän myötä tasapuolisesti jokaiseen tiedonhallintamallin osa-alueeseen liittyen. Puolueellinen tai vajaa näkemys olisi luultavasti saavutettu, jos kaikki haastateltavat olisivat, esimerkiksi dataratkaisujen kehityspuolelta, eikä kenelläkään olisi syvempää kokemusta, esimerkiksi sääntelevien tahojen kanssa toimimisesta. Tutkimuksen luotettavuutta lisää se, että monilla haastateltavilla, niin analytiikan kehityspuolella kuin projektien hallinnointi puolella, on suhteellisen pitkä kokemus, jolloin he ovat nähneet uusien teknologioiden ja regulaatioiden vaikutuksen alalla. Tiedot rooleista ja kokemuksesta on avattu taulukossa 8. Tutkimuksen luotettavuutta vähentää se, että haastateltavien määrä on pieni. Tutkimuksen luotettavuutta olisi parantanut se, että olisi haastateltu, esimerkiksi

terveydenhuollon organisaation henkilöitä aiheeseen liittyen, jolloin olisi saatu parempi näkemys asiakkaan näkökulmasta.

6.4 Jatkotutkimusehdotukset

Tämä tutkimus vastaa hyvin tutkimusaukkoon sekä tutkimuskysymyksiin, jotka määriteltiin luvussa 1.2. Teoriaosuudessa määritellyn tiedonhallintamallin viitekehyksen testi empiriaosuudessa osoittautui myös onnistuneeksi. Asiakkaiden vaatimusten täyttäminen on valmiiksi jo vaikeaa terveysteknologialalla, mutta tässä tutkimuksessa määritelty tiedonhallintamalli auttaa jäsentämään kokonaisuuden ja osoittamaan juuri niihin kohtiin, joihin resurssit kannattaa kohdistaa jo ennalta. Esimerkiksi tietojen varastointiin tarkoitetun palvelimen hankinta EU:n alueelta helpottaa merkittävästi sidosryhmien ja GDPR:n vaatimusten noudattamista, mikä nopeuttaa vaatimustenmukaisuuteen kuluvaa raskasta työtä ja vapauttaa aikaa pk-yritykselle mielekkäämpiin dataprojektivaiheisiin, kuten datan mallinnusvaiheeseen. Big datan hyödyntäminen on ollut trendi jo pitkään pk-yrityksille, mutta sen tehokkaampi ja kokonaisvaltaisempi hyödyntäminen vaatii tiedonhallintamallilta yhä tehokkaampia prosesseja ja menetelmiä. Tämä tutkimus siis onnistui antamaan lisää kontribuutiota tutkimukseen tiedonhallintamallien käytöstä pk-yrityksissä ja nimenomaan terveysteknologia-alalla, koska sitä on aikaisemmin tutkittu todella vähän.

Suurin osa haastateltavista koki viitekehyksen hyödyllisenä terveysteknologia-alan pk-yritykselle. Toisaalta empiriassa tuli myös ilmi se, että viitekehys saattaa olla liian raskas pk-yrityksen resursseihin nähden. Tämä vahvisti teoriaosuudessa todettua asiaa, että tiedonhallintamalli osien, eli pitkälti GDPR vaatimusten, toteutus tehdään pakon edessä ja se koetaan kuormittavana. Yleisesti jatkotutkimusta voitaisiin tehdä siitä, mitä tarkalleen ottaen eri tiedonhallintamallin osa-alueet sisältävät, koska tässä tutkimuksessa pyrittiin selvittämään pitkälti sitä mitkä ovat hyödyllisiä osa-alueita ja hieman kokeilemaan niiden hyödyllisyyttä.

Tutkimusta voisi jatkaa myös niin, että tutkitaan yksinomaan terveydenhuollon organisaatioiden ja regulaatioiden näkökulmaa tiedonhallintamalliin. Tässä tutkimuksessa kävi ilmi, että terveydenhuollon organisaatioiden tiedonhallintamallit eivät ole pk-yrityksien mielestä vielä tarpeeksi kypsällä tasolla. Hyvä asiakkaan tiedonhallintamalli auttaisi pk-yritystä hyödyntämään Big dataa tehokkaammin. Tässä tapauksessa tietojen laatu olisi valmiiksi paremmalla tasolla ja kommunikointi

tietojoukkoihin liittyen sujuisi helpommin, koska asiakkaan tietämys omista tietojoukoista olisi parempi. Empiriaosiossa paljastui jo joitakin näkemyksiä sidosryhmien osalta, kuten se, että tiedonhallintamallin käyttöönotto lisää huomattavasti työtä terveydenhuollon organisaatiossa. Empiriassa ilmeni myös kokemuksia siitä, että sääntelyltä tahoiltakaan ei saa aina selkeää vastausta vaatimuksiin ja että vaatimusten ymmärrys terveydenhuollon organisaatiossa ei ole välttämättä kovin hyvä. Olisi siis mielenkiintoista tutkia tiedonhallintamallin vaikutusta data-analytiikkaprojekteihin terveydenhuollon organisaatioiden ja päättävien tahojen näkökulmasta.

Jatkotutkimuksena voitaisiin tehdä myös tapaustutkimus tiedonhallintamallin vaikutuksesta dataprojekteihin. Tällainen tapaustutkimus antaisi hyvää tietoa siitä, onko tiedonhallintamallista hyötyä CRISP-DM-prosessin eri vaiheisiin terveysteknologia-alalla ja antaisi varmasti mielenkiintoisia havaintoja tiedonhallintamallin käytöstä ja haasteista eri osa-alueissa, kun prosessia seurataan alusta loppuun.

Lähteet

Abraham, R. – Schneider, J. – Vom Brocke, J. (2019). Data governance: A conceptual framework, structured review, and research agenda. International Journal of Information Management, 49, 424-438.

Al-Badi, A. – Tarhini, A. – Khan, A. I. (2018). Exploring big data governance frameworks. Procedia computer science, 141, 271-277.

Alhassan, I. – Sammon, D. – Daly, M. (2016). Data governance activities: an analysis of the literature. Journal of Decision Systems, 25(sup1), 64-75

Annoni, A. – Benczur, P. – Bertoldi, P. – Delipetrev, B. – De Prato, G. – Feijoo, C. ... – Junklewitz, H. (2018). Artificial intelligence: A european perspective.

Begg, C. – Caira, T. (2011). Data Governance in Practice:: The SME Quandary Reflections on the Reality of Data Governance in the Small to Medium Enterprise (SME) Sector. In The European conference on information systems management (p. 75). Academic Conferences International Limited.

Begg, C. – Caira, T. (2012). Exploring the SME Quandary: Data Governance in Practise in the Small to Medium-Sized Enterprise Sector. Electronic Journal of Information Systems Evaluation, 15(1), pp3-13.

Berthold, Borgelt, C. – Höppner, F. – Klawonn, F. (2010). Guide to Intelligent Data Analysis : How to Intelligently Make Sense of Real Data. Springer London. https://doi.org/10.1007/978-1-84882-260-3.

Bäck, A. – Keränen, J. (2017). Anonymisointipalvelut. Tarve ja toteutusvaihtoehdot. Liikenne- ja viestintäministeriö.

Cheatham, B. – Javanmardian, K. – Samandari, H. (2019). Confronting the risks of artificial intelligence. McKinsey Quarterly, 2, 38.

Elliott, V. (2018). Thinking about the coding process in qualitative data analysis. The Qualitative Report, 23(11), 2850-2861.

Eriksson, P. – Kovalainen, A. (2008). Qualitative Methods in Business Research. In Qualitative Methods in Business Research (pp. xii–xii). SAGE Publications. https://doi.org/10.4135/9780857028044

Euroopan parlamentti ja neuvosto. (2016). Asetus luonnollisten henkilöiden suojelusta henkilötietojen käsittelyssä sekä näiden tietojen vapaasta liikkuvuudesta ja direktiivin 95/46/EY kumoamisesta (yleinen tietosuoja-asetus), 2016: < https://eur-lex.europa.eu/legal-content/FI/TXT/PDF/?uri=CELEX:32016R0679&from=FI>, haettu 22.10.2022

Gerke, S. – Minssen, T. – Cohen, G. (2020). Ethical and legal challenges of artificial intelligence-driven healthcare. In Artificial intelligence in healthcare (pp. 295-336). Academic Press.

Germann, S. – Jasper, U. (2020). Realising the benefits of data driven digitalisation without ignoring the risks: health data governance for health and human rights. Mhealth, 6.

Grönfors, M. (2011). Laadullisen tutkimuksen kenttätyömenetelmät. SoFia-Sosiologi-Filosofiapu Vilkka.

Hovi, A. (2020). Data-alan termien selitykset ja kuvaukset: < https://www.arihovi.com/materiaalit/datapedia-data-alan-termit-avattuna/>, haettu 14.11.2022.

Hripcsak, G. – Bloomrosen, M. – FlatelyBrennan, P. – Chute, C. G. – Cimino, J. – Detmer, D. E. ... – Wilcox, A. B. (2014). Health data use, stewardship, and governance: ongoing gaps and challenges: a report from AMIA's 2012 Health Policy Meeting. Journal of the American Medical Informatics Association, 21(2), 204-211.

Huber, S. – Wiemer, H. – Schneider, D. – & Ihlenfeldt, S. (2019). DMME: Data mining methodology for engineering applications–a holistic extension to the CRISP-DM model. Procedia Cirp, 79, 403-408.

IBM. (2021). Structured vs. Unstructured Data: What's the Difference: <https://www.ibm.com/cloud/blog/structured-vs-unstructured-data>, haettu 1.6.2022.

Janssen, M. – Brous, P. – Estevez, E. – Barbosa, L. S. – Janowski, T. (2020). Data governance: Organizing data for trustworthy Artificial Intelligence. Government Information Quarterly, 37(3), 101493.

Kaplan, B. – Maxwell, J. A. (2005). Qualitative research methods for evaluating computer information systems. In Evaluating the organizational impact of healthcare information systems (pp. 30-55). Springer, New York, NY.

Kaushal, A. – Altman, R. – Langlotz, C. (2020). Health care AI systems are biased. Scientific American, 11, 17.

Khatri, V. – Brown, C. V. (2010). Designing data governance. Communications of the ACM, 53(1), 148-152.

Kim, H. Y. – Cho, J. S. (2018). Data governance framework for big data implementation with NPS Case Analysis in Korea. Journal of Business and Retail Management Research, 12(3).

Kroll, J. A. (2018). Data science data governance [AI ethics]. IEEE Security & Privacy, 16(6), 61-70.

Lind, E. – Glas, S. (2022). DATA MINING IN PRACTICE: An application of the CRISP-DM framework in healthcare.

Nummenmaa, L. – Pulkkinen, P. – Holopainen, M. (2016). Tilastollisten menetelmien perusteet (1.-2. p.). Helsinki: Sanoma Pro.

Muehlematter, U. J. – Daniore, P. – Vokinger, K. N. (2021). Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. The Lancet Digital Health, 3(3), e195-e203.

Niemi, E. (2011, August). Designing a data governance framework. In Proceedings of the IRIS Conference, At Oslo, Norway (Vol. 14).

Nwabude, C. – Begg, C. – McRobbie, G. (2014). Data governance in small businesses-why small business framework should be different. International Proceedings of Economics Development and Research, 82, 101-107.

Okoro, R. (2021). Proposed Data Governance Framework for Small and Medium Scale Enterprises (SMES). Minnesota State University, Mankato.

Otto, B. (2011). A morphology of the organisation of data governance.

Panian, Z. (2010). Some practical experiences in data governance. World Academy of Science, Engineering and Technology, 62(1), 939-946.

Rahm, E. – Do, H. H. (2000). Data cleaning: Problems and current approaches. IEEE Data Eng. Bull., 23(4), 3-13.

Rodrigues, I. (2020). CRISP-DM methodology leader in data mining and big data: < https://towardsdatascience.com/crisp-dm-methodology-leader-in-data-mining-and-big-data-467efd3d3781>, haettu 25.10.2022.

Safran, C. – Bloomrosen, M. – Hammond, W. E. – Labkoff, S. – Markel-Fox, S. – Tang, P. C. – Detmer, D. E. (2007). Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. Journal of the American Medical Informatics Association, 14(1), 1-9.

Saltz, J. – Hotz, N. – Wild, D. – Stirling, K. (2018). Exploring project management methodologies used within data science teams.

Schröer, C. – Kruse, F. – Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. Procedia Computer Science, 181, 526-534.

Shah, A. – AlSousi, A. (2021). DATA GOVERNANCE: A CONCEPTUAL FRAMEWORK FOR SME. The Middle East International Journal for Social Sciences (MEIJSS).

Sirur, S. – Nurse, J. R. – Webb, H. (2018, January). Are we there yet? Understanding the challenges faced in complying with the General Data Protection Regulation (GDPR). In Proceedings of the 2nd International Workshop on Multimedia Privacy and Security (pp. 88-95).

Taanila, A. (2019). Ruutu- ja janakuvio: <https://tilastoapu.wordpress.com/2013/05/02/laatikkokaavio/>, haettu 19.11.2022.

Tevameri, T. (2018). Syksyn 2018 toimialojen näkymät: Pk-yritysten asemaan ja markkinoiden muutoksiin on kiinnitettävä huomiota sote-alalla.

Tudorica, B. G. – Bucur, C. (2011, June). A comparison between several NoSQL databases with comments and notes. In 2011 RoEduNet international conference 10th edition: Networking in education and research (pp. 1-5). IEEE.

Tuomi, J. – Sarajärvi, A. (2018). Laadullinen tutkimus ja sisällönanalyysi (Uudistettu laitos.). Helsinki: Kustannusosakeyhtiö Tammi.

Weber, K. – Otto, B. – Österle, H. (2009). One size does not fit all---a contingency approach to data governance. Journal of Data and Information Quality (JDIQ), 1(1), 1-27.

Winter, J. S. (2021). AI in healthcare: Data governance challenges. Journal of hospital management and health policy, 5(8).

Winter, J. S. – Davidson, E. (2019). Big data governance of personal health information and challenges to contextual integrity. The Information Society, 35(1), 36-51.

Liitteet

Liite 1. Haastattelukysymykset

Johdanto

1 Kerro lyhyesti taustasi terveysteknologia-alalla?

2 Miten määrittelisit tiedonhallintamallin (data governance) lyhyesti?

3 Kerro lyhyesti kokemuksesi dataprojekteista?

Sidosryhmien vaatimukset

4 Millainen tietämys sidosryhmillä on teknisistä ratkaisuista?

   a. Jos sidosryhmillä ei ole tietämystä niin mitkä tekijät ovat olleet esteenä tietämykselle?

5 Millaisia vaikeuksia yrityksellä on ollut projektin ymmärtämisessä?

   a. Miksi vaikeuksia on ollut?

6 Minkä tyyppisiä vaatimuksia dataprojektien sidosryhmillä on?

   a. Tulevatko vaatimukset suoraan GDPR:stä?

   b. Tuleeko vaatimuksia muista säädöksistä tai standardeista?

Tietojen läpinäkyvyys

7 Millaisia vaatimuksia terveydenhuollon organisaatiolla on dataprojektin dokumentoinnille?

8 Dokumentoidaanko tietojoukkoihin tehtävät muutokset erikseen muistiin?

9 Millä tarkkuudella algoritmien toiminta on dokumentoitu?

Tietojen laatu

10 Miten tietojoukkojen laatua parannetaan?

11 Onko data heti käyttökelpoista?

12 Jos dataa joudutaan muokkaamaan niin miksi? Ovatko tiedot epäjohdonmukaisia, virheellisiä tai puolueellisia?

13 Ovatko luovutettavat tiedot yleensä ajankohtaisia?

Tietojen laajuus

14 Onko data yhdistelty eri tietolähteistä jo valmiiksi?

   a. Jos ei niin joutuuko yritys tekemään tämän itse?

15 Tarjoaako terveydenhuollon organisaatio dataprojektin tavoitteisiin nähden riittävän tietojoukon?

Tietovastaavan määrittäminen

16 Millä perusteilla yrityksessä valitaan tietovastaavaksi?

17 Millaisia vaatimuksia sidosryhmillä on tietovastaavan kompetenssille?

Tietojen prosessointi

18 Millaisia vaatimuksia datan prosessointimetodien valintaan liittyy?

19 Millä tavoin varmistetaan tietojen prosessointitapojen riittävyys?

20 Millä tavoin datan prosessoinnin reiluus ja asianmukaisuus varmistetaan yksilöön nähden?

Liite 2. Aineistonhallintasuunnitelma

Tutkimusaineisto

Aineistotyyppi	Sisältää henkilötietoja*	Tuotan aineiston itse	Joku muu on tuottanut aineiston	Muuta huomioitavaa
Aineistotyyppi 1: *Haastattelut*	x	x		
Aineistotyyppi 2: Nauhoitukset	x	x		
Aineistotyyppi 3: Omat muistiinpanot		x		

Henkilötietojen käsittely tutkimuksessa

Mikäli aineistosi sisältää henkilötietoja, olet velvoitettu noudattamaan EU:n tietosuoja-asetusta (GDPR) sekä Suomen tietosuojalakia. Henkilötietoja sisältävän aineiston osalta sinun tulee laatia tutkittavillesi tietosuojailmoitus sekä selvittää, kuka toimii aineiston osalta rekisterinpitäjänä.

Laadin tutkittavilleni tietosuojailmoituksen** ja toimitan sen heille ennen aineiston keruuta ☐

Henkilötietojen osalta rekisterinpitäjänä** toimii opiskelija ☐ yliopisto ☐

Aineistoni ei sisällä henkilötietoja ☒

**Lisätietoja yliopiston intranetin Tietosuojaohjeita opinnäytetyöhön -sivulta

Aineiston käyttöön liittyvät luvat ja oikeudet

Selvitä mitä lupia ja oikeuksia aineistojen käyttöön liittyy. Ole tarvittaessa yhteydessä opinnäytteesi ohjaajaan. Kuvaile jokaisen aineistotyypin osalta niiden käyttöön liittyvät luvat ja oikeudet, voit tarvittaessa lisätä aineistotyyppejä listaukseen.

Itse tuotettu aineisto

Saatat tarvita erillisiä lupia keräämäsi tai tuottamasi aineiston käyttöön sekä tutkimuksessa että tulosten julkaisemisessa. Mikäli olet arkistoimassa aineistoasi, pyydä

tutkittavilta tarvittavat luvat aineiston arkistoimiseen ja jatkokäyttöön. Selvitä myös, vaatiiko valitsemasi arkisto kirjallisia lupia tutkittavilta.

Tarvittavat luvat ja niiden hankkiminen

Aineistotyyppi 1, 2 ja 3: Haastattelut, nauhoitukset ja muistiinpanot

Haastateltaville tuodaan esille seuraavat asiat ennen haastatteluja:

- Haastatteluihin osallistuminen on vapaaehtoista
- Haastattelut nauhoitetaan tai tallennetaan ja litteroidaan tulosten analysoimiseksi
- Haastateltavilta ei kerätä henkilötietoja tai mitään tietoja, joista heidät voisi tunnistaa
- Nauhoitukset, litteroidut haastattelut ja muistiinpanot säilytetään Yliopiston ylläpitämässä Seafile-pilvipalvelussa sekä omalla tietokoneella, jossa on virustorjuntaohjelma
- Kerättyä dataa aineistoa käytetään vain tutkimuksen tekemiseen

Aineiston säilyttäminen tutkimuksen aikana

Missä säilytät aineistoasi tutkimuksen aikana?

Yliopiston verkkokansiossa ☐

Yliopiston tarjoamassa Seafile-pilvipalvelussa ☒

Jossakin muualla, missä? ☒

Aineistoa säilytetään myös omalla tietokoneella, joka sisältää virustorjuntaohjelman.

Yliopiston tallennuspalvelut huolehtivat automaattisesti tietoturvasta ja varmuuskopioinnista. Jos valitset tallentamisen muualle kuin yliopiston palveluihin, kuvaa, miten huolehdit tietoturvasta ja varmuuskopioinnista. Muista varmistaa, mihin tallennat aineiston aina sitä muokattuasi.

Aineiston dokumentointi ja metadata

Miten kuvailet aineistosi niin, että ulkopuolinenkin ymmärtää, millaista aineisto on? Miten itse tarpeen tullen palautat vuosien kuluttua mieleesi, mistä aineistosi koostuu?

Aineiston dokumentointi

Pystytkö kertomaan, mitä aineistollesi on tapahtunut tutkimuksen teon aikana? Aineiston dokumentointi on keskeisessä osassa aineistoon tehtyjen muutosten jäljittämisessä.

Käytän aineiston dokumentointiin

tutkimuspäiväkirjaa ☐

erillistä dokumenttia, johon kirjaan aineiston pääasiat, kuten tehdyt muutokset, analyysin vaiheet sekä esim. muuttujien merkitykset ☒

aineiston mukana kulkevaa readme-tiedostoa, jossa kuvataan aineiston pääasiat ☐

jotain muuta, mitä? ☐

Aineiston järjestys ja eheys

Miten pidät aineistosi järjestyksessä ja ehyenä, ja vältät sen tahattomat muutokset?

Säilytän alkuperäisen aineiston erillään tutkimuksenteon aikana käyttämästäni aineistosta, jotta voin palata alkuperäiseen, jos tarvetta ilmenee. ☒

Versionhallinta: mietin jo ennen tutkimuksenteon alkua, miten tulen nimeämään eri aineistoversiot ja noudan sitä systemaattisesti ☒

Tiedostan jo tutkimuksen alussa aineistoni elinkaaren, ja varaudun tilanteisiin, joissa data saattaa huomaamatta muuttua, kuten esim. nauhoitus, litterointi, konversio toiseen tiedostomuotoon, tallentaminen jne. ☒

Metadata

Metadata on kuvaus aineistostasi. Metadatan perusteella henkilö, joka ei tunne aineistoasi, ymmärtää, millaista aineistosi on. Metadataa voi olla mm. tiedoston nimi, sijainti, koko ja tieto aineiston tuottajasta. Tarvitsetko metadataa?

Tallennan aineistoni arkistoon tai tietopankkiin, joka huolehtii metadatasta puolestani. ☐

Minun pitää luoda metadata, koska arkisto, johon tallennan aineiston edellyttää sitä. ☐

En tallenna aineistoani julkiseen arkistoon, enkä tarvitse metadataa. ☒

Aineisto tutkimuksen valmistuttua

Olet vastuussa aineistostasi myös tutkimuksen valmistumisen jälkeen. Varmista, että käsittelet sitä tekemiesi sopimusten mukaisesti. Yliopiston suosittelema säilytysaika on viisi vuotta, poikkeuksena kuitenkin lääketieteen alan aineistot, joiden säilytysaika on 15 vuotta. Henkilötietoja voi säilyttää vain sen aikaa, kun tarve on. Jos olet sitoutunut tuhoamaan aineiston määräajan päätyttyä, sinun on huolehdittava siitä, vaikka et olisi enää opiskelija. Myös yliopiston tallennusratkaisuja käytettäessä aineiston tuhoaminen on sinun vastuullasi.

Mitä aineistollesi tapahtuu, kun tutkimus valmistuu?

Tutkimusdata säilytetään Turun yliopiston suosituksen mukaan viisi vuotta.

Jos säilytät dataa, kuvaa, missä: Omalla tietokoneella

Aineistonhallintasuunnitelma kannattaa pitää ajan tasalla läpi tutkimuksen.

Lisätietoja Turun yliopiston kirjaston laatimasta Opiskelijan aineistonhallintaoppaasta