Nelli Kiianmaa

KLIINISEN BIG DATAN LAATUONGELMAT JA NIIDEN SYYT TIETOALLASYMPÄRISTÖSSÄ

JYVÄSKYLÄN YLIOPISTO
INFORMAATIOTEKNOLOGIAN TIEDEKUNTA
2020

TIIVISTELMÄ

Kiianmaa, Nelli
Kliinisen big datan laatuongelmat ja niiden syyt tietoallasympäristössä
Jyväskylä: Jyväskylän yliopisto, 2020, 103 s.
Tietojärjestelmätiede, pro gradu -tutkielma
Ohjaajat: Koskelainen, Tiina; Seppänen, Ville; Taipalus, Toni

Terveydenhuollon kliinistä tietoa ja big dataa, kuten sairauskertomustietoa, hyödynnetään enenevässä määrin toissijaisiin tarkoituksiin, kuten tutkimukseen ja tiedolla johtamiseen. Tietolähteet ovat hyvin moninaisia ja tiedon laatu alhainen, mikä hankaloittaa tiedon käyttöä. Laatuongelmiin on big dataa käsittelevässä kirjallisuudessa kiinnitetty verrattain vähän huomiota.

Tässä tutkimuksessa tutkittiin kliinisen big datan laatuongelmia, niiden syitä ja niihin kehitettyjä ratkaisuja tutkimuksessa ja tiedolla johtamisessa. Näkökulma oli sosiotekninen. Tutkimus toteutettiin laadullisena tapaustutkimuksena Varsinais-Suomen sairaanhoitopiirin tietoallasympäristössä ja urologian tietoallashankkeessa. Aineisto koostui puolistrukturoiduista haastatteluista ja julkisista dokumenteista. Analyysimenetelminä käytettiin aineisto- ja teorialähtöistä sisällönanalyysiä sekä visualisointia.

Tapauskontekstissa tiedon laatuongelmia syntyy kaikissa vaiheissa potilastiedon kirjaamisesta sen pohjalta tehtyihin johtopäätöksiin asti. Laatuongelmien syyt ovat moninaisia ja kytköksissä toisiinsa. Tietoaltaan potilastietojen relevanssi ja arvo toissijaisessa käytössä on lähtökohtaisesti heikko. Syynä on potilastietojen kirjaamisen muoto ja tapa, erityisesti rakenteisen tiedon puute. Rakenteisen tiedon puuttuessa on käytettävä sairauskertomustekstiä, jonka hyödyntäminen on vaativaa. Tiedon varastointi- ja jalostusvaiheessa tiedon laatuongelmia aiheuttaa tiedon sirpaleisuus, viiteavainten ja metatiedon puute sekä monipolvinen, virhealtis jalostusprosessi. Ilman riittäviä osaamis- ja teknologiaresursseja tietoallastiedon tehokas hyödyntäminen ei ole mahdollista. Urologian tietoallashankkeessa tiedon laatuongelmia pyrittiinkin ratkaisemaan erityisesti panostamalla kliinikoiden ja it-asiantuntijoiden yhteiseen, pitkäjänteiseen kehitystyöhön.

Tutkimustulokset auttavat ymmärtämään, mitkä ovat keskeisiä kehityskohteita, kun kliinisestä tiedosta pyritään jalostamaan arvoa tietoallasympäristössä.

Asiasanat: terveydenhuolto, big data, toissijainen käyttö, toisiokäyttö, tiedolla johtaminen, tiedon laatu, tiedon laatuongelmat

ABSTRACT

Kiianmaa, Nelli
Quality problems of clinical big data and their causes in a data lake environment
Jyväskylä: University of Jyväskylä, 2020, 103 pp.
Information Systems, Master's Thesis
Supervisors: Koskelainen, Tiina; Seppänen, Ville; Taipalus, Toni

Healthcare clinical data and big data, such as electronic health record, are increasingly being utilised for secondary purposes, such as research and knowledge management. The data sources are very diverse, and of low quality, creating challenges for their use. Relatively little attention has been paid to quality problems in the big data literature.

This study examined the quality problems of clinical big data, their causes, and the solutions developed in research and knowledge management. The perspective was socio technical. The study was carried out as a qualitative case study in the data lake environment of the Hospital District of Southwest Finland and in their urology data lake project. The research material consisted of semi-structured interviews and public documents. Conventional and directed content analysis and visualisation were used as analysis.

In the case context, data quality problems arise at all stages from the recording of patient data to the conclusions derived. The causes of quality problems are manifold and interconnected. The relevance and value of the data lake patient data in secondary use is weak per se. This is due to the form and manner of recording patient data, especially the lack of structured information. In the absence of structured information, narrative text must be used, the utilization of which is demanding. In the data storage and processing phase, data quality problems are caused by data fragmentation, lack of reference keys and metadata, and a multi-phase, error-prone processing process. Without sufficient know-how and technology resources, effective utilisation of data lake information is not possible. The urology data lake project sought to solve data quality problems, especially by investing in the joint, long-term development work of clinicians and IT experts.

The results help to understand the key areas for development in the pursuit of acquiring value from clinical data in a data lake environment.

Keywords: healthcare, big data, secondary use, reuse, knowledge management, data quality, data quality problems

KUVIOT

TAULUKOT

SISÄLLYS

1 JOHDANTO

Terveydenhuollossa kertyy valtavat määrät tietoa yhä kiihtyvällä vauhdilla (Feldman, Martin, & Skotnes, 2012). Puhutaan niin sanotusta big datasta, jolla tarkoitetaan useimmiten määrällisesti suurta määrää moninaista tietoa, jota luodaan, tallennetaan ja prosessoidaan suurella nopeudella (Mikalef ym., 2018). Toiveet big datasta jalostettavalle arvolle ovat suuret (Baesens ym., 2016), mutta big datan alhainen laatu on kuitenkin haaste sen hyödyntämisessä (esim. Buhl ym., 2013).

Terveydenhuollossa tietoa kerätään kirjaamalla ja tallentamalla tietoa potilaista ja heidän hoidostaan terveydenhuollon tietojärjestelmiin muun muassa lääkärin vastaanotolla, kuvantamisen yksikössä ja laboratoriossa. Terveystietoa syntyy myös muun muassa älylaitteiden, kuten kotihoitosensorien, avulla ja esimerkiksi sosiaalisen median palveluissa. (Mehta & Pandit, 2018.) Tällainen tieto on pitkään ollut terveydenhuollossa toiminnan oheistuotteen asemassa, eikä sitä ole ymmärretty keskeiseksi voimavaraksi (Murdoch & Detsky, 2013). Kiinnostus tiedon hyödyntämiseen on kuitenkin kasvussa. Väestöjen ikääntyminen ja ihmisten muuttuva elämäntyyli lisäävät terveydenhuollon järjestelmiin kohdistuvaa painetta kaikkialla maailmassa (Kankanhalli ym., 2016). Terveydenhuollon menot kasvavat, ja big datan avulla toivotaan saavutettavan huomattavia kustannussäästöjä, laadukkaampaa ja tehokkaampaa hoitoa. (Feldman ym., 2012.)

Terveydenhuollon big datan haasteet ovat erilaisia kuin liiketoiminnan big datan. Terveydenhuollon big data on luonteeltaan moninaista, eri tahojen tietovarastoihin siiloutunutta sekä turvallisuus- ja tietoturvakriittistä. (Jee & Kim, 2013.) Haasteita terveydenhuollon big datan käytölle ovat tiedon laatuun ja sen moninaisuuteen liittyvät ongelmat, tietoturvan kysymykset, tiedon omistajuus, monimutkaiset säädökset, sopivien IT-infrastruktuurien puute, suuret analyyttisten välineiden investointikustannukset ja korkeatasoisen osaamisen puute (Feldman ym., 2012; Mehta & Pandit, 2018).

Edistysaskeleet terveydenhuollon big datan hyödyntämisessä riippuvat paremmista tavoista ottaa käyttöön erilaisia jo olemassa olevia tiedonlähteitä ja tietoaltaita ja toisaalta uudenlaisen tiedon virtoja. (Feldman ym., 2012.) Suomessa edellä mainittuja tiedon sujuvan hyödyntämisen esteitä pyritään poistamaan kansallisin sosiaali- ja terveysalan digitalisaatiohankkein, joita on toteutettu jo useiden vuosien ajan. Tavoitteena on yhtenäistää hajanaisia järjestelmiä, parantaa järjestelmien käytettävyyttä ja tehostaa tiedon hyödyntämistä. Myös lainsäädäntöä on uudistettu. Tulevaisuudennäkymänä on, että yhtenäinen

infostruktuuri ja uudet tietojärjestelmät tulevat mahdollistamaan myös tiedon entistä monipuolisemman ja nopeamman keräämisen ja käytön. (Hyppönen & Ilmarinen, 2016.)

Yleisesti big datan hyödyntämisessä yhdistyvät suuret mahdollisuudet ja riskit (Clarke, 2016). Jo pitkään on tiedetty, että huonolaatuinen tieto aiheuttaa organisaatioille valtavia kustannuksia (Strong, Lee, & Wang, 1997). Tiedon laatu vaikuttaa keskeisesti myös analytiikasta saatuun hyötyyn (Ghasemaghaei & Calic, 2019). Big datan laatu on useimmiten suhteellisen alhainen, ja huonolaatuisen big datan kohdalla riskit voivat olla erityisen suuret, kun sitä käytetään päätöksenteon pohjana. Tämä voi johtaa resurssien jakamiseen väärin ja kokonaisten väestönosien epäoikeudenmukaiseen suosimiseen tai syrjintään. (Clarke, 2016.) Riskit korostuvat terveydenhoidossa, jossa huonot päätökset vaikuttavat paitsi talouteen myös ennen kaikkea ihmisten terveyteen (Wang ym., 2019).

Big datan laatu on siihen liittyvistä riskeistä huolimatta jäänyt tutkimuskirjallisuudessa varjoon (Baesens ym., 2016). Erityisesti empiirinen tutkimus on niukkaa (Galetsi ym., 2020). Lisäksi big datan tutkimus on tähän saakka ollut teknispainotteista ja inhimilliset tiedot ja taidot on suurelta osin unohdettu (Mikalef ym., 2018). Tyypilliset big datan määritelmät hämärtävät tiedon keräämisen, analysoinnin ja käytön organisatoriset käytännöt (Markus & Topi, 2015). Markuksen ja Topin (2015) mukaan olisikin tärkeää tarkastella big dataa sosioteknisestä näkökulmasta, joka huomioi big dataan liittyvät ideat, resurssit ja käytännöt. Tällaisen näkemyksen mukaan big data ei ole vain tietoa vaan sisältää myös kaiken sen, mitä sillä tehdään tai voidaan tehdä sekä sen käyttöä ohjaavat tavoitteet ja arvot (Markus & Topi, 2015). Esimerkiksi big data innovaationa sisältää sekä teknologiat ja välineet että tiedon, taidot, käsitteet, organisaatiot ja muut sosiaaliset ja institutionaaliset kontekstit (Chae, 2019).

Tässä tutkimuksessa pyritään pureutumaan edellä mainittuihin tutkimusaukkoihin, ja tuottamaan empiiristä tutkimustietoa terveydenhuollon big datan laatuongelmista sosioteknisestä näkökulmasta. Sairauskertomustieto eli potilaista ja heidän hoidostaan terveydenhuollossa kerätty kliininen tieto on keskeistä tutkimukseen ja muihin toissijaisiin tarkoituksiin hyödynnettävää tietoa terveydenhuollossa (Weiskopf & Weng, 2013). Sen arvo korostuu, kun siihen yhdistetään muista lähteistä saatua aineistoa, esimerkiksi genomitietoa (Costa, 2014; Dinov, 2016). Terveydenhuollon big datan laatuongelmat ovatkin suurelta osin sairauskertomustiedon laatuongelmia (ks. esim. Hoffman, 2014). Terveydenhuollon big datasta saatavan arvon kannalta sairauskertomuksen ja muun kliinisen tiedon laatuongelmien ratkaiseminen ja tiedon mahdollisimman tehokas hyödyntäminen on tärkeää.

Tutkimuksen tarkoituksena on selvittää laadullisen tapaustutkimuksen keinoin, mitä laatuongelmia suomalaisessa terveydenhuollon big datassa on ja mistä ne aiheutuvat. Kontekstina ja rajauksena on yleisesti Varsinais-Suomen sairaanhoitopiirin (VSSHP) tietoaltaan kliinisen tiedon käyttö tiedolla johtamiseen ja tutkimukseen Turun yliopistollisessa keskussairaalassa (TYKS) sekä erityisesti urologian klinikan tietoallashanke. Hankkeessa on kehitetty lääkärin tueksi potilastiedon visuaalinen aikajananäkymä sekä tiedonkeruuta ja raportointia erityisesti eturauhassyövän hoidon laadun mittaamiseen. Hanke on tietoallashankkeena askel kohti big datan laajempaa hyödyntämistä suomalaisessa terveydenhuollossa.

Tutkimuskysymyksiä on kaksi:

1. Mitä big datan laatuongelmia kontekstissa on ollut ja mistä ne johtuvat?

2. Miten big datan laatuongelmia on pyritty ratkaisemaan urologian tietoallashankkeessa?

Tutkimuksen tarkoituksena on tuottaa ajankohtaista tietoa siitä, mikä estää tai vaikeuttaa big datan hyödyntämistä tapauskontekstissa it-asiantuntijoiden ja kliinikoiden näkökulmasta. Tavoitteena on, että terveydenhuollon toimijat voisivat hyödyntää tutkimuksen tuloksia toiminnassaan ja kehittämistyössään. Tutkimuksessa käsitellään kokemuksia organisaation sisäisestä tiedon hyödyntämisestä, eikä se ota kantaa siihen, miten tiedon laatuongelmat näyttäytyvät organisaation ulkoisten käyttäjien näkökulmasta. Tiedon laatuongelmia ei myöskään käsitellä teknisellä tasolla tai pyritä kehittämään niihin konkreettisia ratkaisuja.

Tutkielman rakenne on seuraava. Ensimmäinen osa, luvut kahdesta kolmeen, muodostavat kirjallisuuskatsauksen. Luvussa kaksi keskustellaan big datasta tietona, big data -teknologioista ja big datan sovellutuksista sekä luodaan katsaus tiedon ja big datan laatuun ja laatuongelmiin. Luku kolme keskittyy big dataan ja tiedon laatuun ja laatuongelmiin terveydenhuollossa ja erityisesti sairauskertomustiedon toissijaisessa käytössä. Empiirinen osuus alkaa menetelmäluvulla (luku neljä), jossa esitellään ja perustellaan käytetyt aineistonkeruu- ja analyysimenetelmät sekä kuvataan tutkimuksen konteksti ja empiirisen osan toteutus. Luvussa viisi esitellään analyysin tulokset, jota seuraa tulosten tarkastelu ja pohdinta luvussa kuusi.

2 BIG DATA LISÄÄ TIEDON LAATUONGELMIA

Big dataa -ilmiönä ei olisi ilman 1990-luvun alussa alkanutta voimakasta digitalisaatiota, analogisen tiedon muuntamista digitaaliseen muotoon (De Mauro, Greco, & Grimaldi, 2016). Kun puhutaan big datasta, puhutaan määrältään suuresta, moninaisesta ja nopeasti kertyvästä digitaalisesta tiedosta ja sen jalostamisesta arvoksi (Mikalef ym., 2018). Tässä luvussa keskitytään ensimmäiseksi big datan määritelmiin ja big dataan tietona. Lisäksi käsitellään big datan ominaisuuksien käsittelyyn kehitettyjä big data -teknologioita sekä big datan sovellutuksina big data -analytiikkaa ja päätöksentekoa. Samalla puhutaan siitä, mitä big datan hyödyntäminen vaatii organisaatioilta. Big datan laatu on yleisesti heikko, mikä on esteenä sen tehokkaalle jalostamiselle arvoksi (Baesens ym., 2016). Luvun toisena aihepiirinä onkin tiedon laatu: tiedon laadun määritelmät, tiedon laatu-ulottuvuudet sekä laatuongelmat ja niiden syyt. Lopuksi keskustellaan tiedon laadusta big data -kontekstissa.

Kirjallisuushaun tavoitteena oli löytää mahdollisimman laadukkaita, relevantteja ja tuoreita lähteitä. Hakuun käytettiin Google Scholar -hakukonetta hakusanoin *big data, data quality, information quality, data quality issues* ja *data quality problems.* Kirjallisuutta löydettiin myös lumipallotekniikan avulla. Lisäksi on käytetty joitakin suomenkielisiä lähteitä. Suomeksi big datasta puhutaan usein termillä massadata (esim. Jalonen, 2015) viitaten sen suureen määrään. Tässä tutkielmassa käytetään kuitenkin alkuperäistä englanninkielistä ilmaisua, sillä se on vakiintunut myös suomenkieliseen käyttöön ja ilmaisee laajasti big dataan liittyviä ideoita.

2.1 Big data

2.1.1 Big datan määritelmä

Big dataa koskevaa kirjallisuus on nuorta ja hajanaista (De Mauro ym., 2016; Kitchin & McArdle, 2016). Vaikka sitä on noin vuodesta 2000 (Chen, Hsinchun & Chiang, 2012; Günther ym., 2017), laajemmalti termi yleistyi vasta vuodesta 2011 (Gandomi & Haider, 2015). Useimmat big datan määritelmät nojaavat Laneyn (2001; Gandomi & Haider (2015) mukaan) (Gartner) kuvaukseen kolmiulotteisesta tiedosta, jota luonnehtii suuri määrä (vo-

lume), moninaisuus (variety) sekä luomisen, tallennuksen ja prosessoinnin suuri nopeus (velocity) (”kolme V:tä”) (Mikalef ym., 2018; Mauro ym., 2016; Gandomi & Haider, 2015). Lisäksi määritelmään liitetään lisäksi alhainen todenmukaisuus (veracity) ja suuri arvo (value) (”viisi V:tä”) (esim. Baesens ym., 2016).

Varhaiset big datan määritelmät ovat peräisin suurista teknologiakonsultointiyhtiöistä, kuten Gartner ja McKinsey, ja niissä painottuu teknologisten kynnysten ylittäminen. Manyikan ym. (2011) (McKinsey) mukaan big data “viittaa aineistoihin, joiden koko on liian suuri, jotta niitä voisi tallentaa, varastoida, hallinnoida ja analysoida tyypillisin tietokantasovelluksin”. Beyer ja Laney (2012, Mikalef ym., 2018 mukaan**)** (Gartner) puolestaan määrittelevät big datan ”suureksi määräksi hyvin nopeita ja/tai moninaisia informaatiovoimavaroja, jotka vaativat uusia prosessoinnin muotoja parantaakseen päätöksentekoa, ymmärryksen syventämistä sekä prosessien automaatiota.” Myöhemminkin konsultointi- tai teknologiayhtiöillä on ollut suuri vaikutus big datan ominaisuuksien määrittelyssä (ks. esim. Gandomi & Haider, 2015).

De Mauron ym. (2016) mukaan big data -termillä kuvataan tieteellisessä kirjallisuudessa monia eritasoisia asioita: sen edustamaa sosiaalista ilmiötä, informaatiota voimavarana, aineistoja, tallennusteknologioita, analyysitekniikoita, prosesseja ja infrastruktuureita. He jakavat erilaiset big datan määritelmät sen mukaan, onko niissä kuvattu big dataa tietona, teknologioina, menetelminä vai niiden vaikutusten kautta, joita sillä on yhteiskunnan eri alueilla. Laajan kirjallisuuskatsauksen perusteella he määrittelevät big datan ”informaatiovoimavaraksi, jota luonnehtii suuri määrä, nopeus ja moninaisuus, ja joka vaatii erityisiä teknologioita ja analyyttisiä menetelmiä informaation muuntamiseksi arvoksi”. (De Mauro ym., 2016.)

Big datan määrä viittaa usein puhtaasti aineiston kokoon eli muuttujien ja havaintojen määrään (George ym., 2016). Big datalla on kokoa teratavuista tai petatavuista eksatavuihin (Abbasi, Sarker, & Chiang, 2016; Chen et al., 2012). Tiedon suuri määrä on kuitenkin kontekstisidonnainen, ja sen rajat ovat taipuvaisia muuttumaan ajan myötä (Manyika ym., 2011; Gandomi & Haider, 2015).

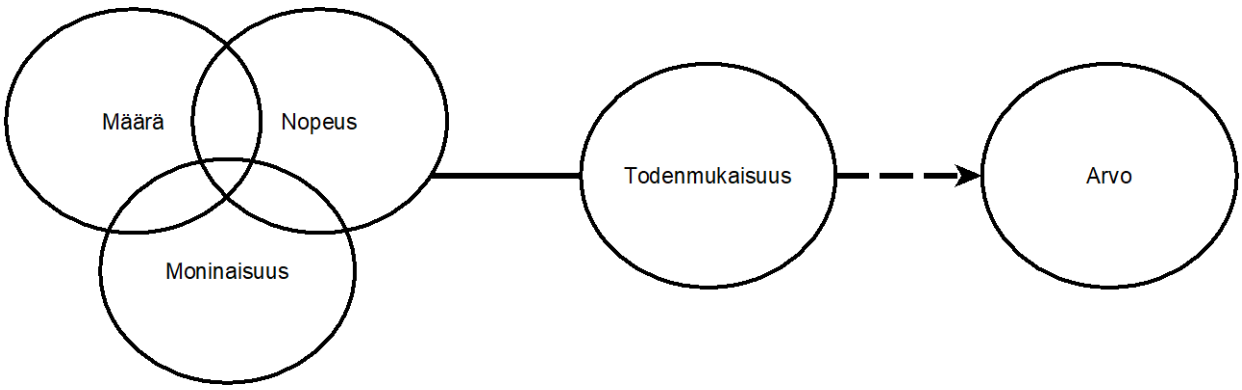
Big datan moninaisuus perustuu vaihtelevaan sisältöön sekä vaihteleviin formaatteihin ja kommunikaation tyyppeihin (Constantiou & Kallinikos, 2015; Davis, 2014), ja viittaa big data -aineiston rakenteelliseen heterogeenisuuteen (Gandomi & Haider, 2015). Tieto voi olla transaktiodataa, käyttäjien luomaa tekstiä, kuvia, videoita, sosiaalisen verkoston tietoa, sensoridataa, verkon ja mobiilia ja spatiotemporaalista tietoa (Chen, Hsinchun, Roger H. L. Chiang, 2012; McAfee & Brynjolfsson, 2012). Se voi olla rakenteista, puolirakenteista, heikosti rakenteista tai rakenteetonta (Gandomi & Haider, 2015; Batini, Palmonari, & Viscusi, 2014). Vain erittäin pieni osa big datasta on rakenteista tietoa, jota voidaan tallentaa perinteisiin relaatiotietokantoihin (Davenport, Barth & Bean, 2012). Aineiston heterogeenisuuden lisäksi voidaan viitata myös tiedon lähteiden moninaisuuteen (Clarke, 2016; Yoo, 2015; Baesens ym., 2016). Big datan tärkeimpiä lähteitä ovat organisaatioiden suuret tietojärjestelmät, sosiaalinen media, mobiililaitteet, esineiden internet sekä avoimet ja julkiset tiedonlähteet (Baesens ym., 2016). Moninaisuuden ohella käsitteellä kompleksisuus voidaan viitata eri lähteistä saadun tiedon yhdistämiseen, puhdistamiseen ja muokkaamiseen (Gandomi & Haider, 2015).

Nopeus viittaa tiedon koko elinkaareen, joka on big datalla usein lyhyt. Se tarkoittaa tiedon luomisen, keräämisen, varastoimisen, prosessoinnin, päivittämisen ja analysoimisen sekä vanhenemisen vauhtia (Abbasi, Sarker, & Chiang, 2016; Davis, 2014; George, ym.,

2016; Khine & Wang, 2018). Digitaalisten laitteiden, kuten älypuhelinten ja sensoreiden, lisääntyminen on johtanut ennennäkemättömään tiedon luomistahtiin (Gandomi & Haider, 2015). Tietoa syntyy lähes tai täysin reaaliajassa (Kitchin, 2013). Big datan tietovirrat ovat paitsi määrältään suuria usein myös vaihtelevia, jolloin ajoittain virrassa on piikkejä ja aallonpohjia (Gandomi & Haider, 2015). Tämä on haaste muun muassa big data -järjestelmien suorituskyvyn arvioinnille (Xiong ym., 2013) sekä big datan jalostamisen kannalta (Abbasi ym., 2016).

Monet tutkijat sisällyttävät todenmukaisuuden big datan määritelmään (Gandomi & Haider, 2015). Big data -kontekstissa todenmukaisuus liittyy kaikkeen sellaiseen, mikä voi vähentää tiedon paikkansapitävyyttä tai tekee päättelystä tiedon perusteella epävarmaa, kuten epäyhdenmukaisuus, puuttuva tieto, epäselvyys, vilppi ja latenssi. Big data sisältää jo määritelmällisestikin laadun heikkouksia. Tiedon suuri määrä voi peittää tiedon huonon laadun, nopeus moninkertaistaa sen ja moninaisuus aiheuttaa tiedon ja kontekstin välistä epäselvyyttä. (Sukumar, Natarajan, & Ferrell, 2015.) Cappiello, Samá, & Vitali (2018) viittaavat todenmukaisuudella sellaisiin tiedon puutteisiin tai virheisiin, jotka huonontavat sen käytettävyyttä. Usein viitataan myös erilaisten tietolähteiden uskottavuuden ja luotettavuuden vaihteluun (Abbasi ym., 2016). Todenmukaisuuteen voidaan sisällyttää myös muita tiedon laatuominaisuuksia ja esimerkiksi tiedonhallinnan tai tietoturvan vaatimuksia (ks. esim. Demchenko ym., 2013; Wang ym., 2019).

Big dataan liittyvä suuri arvolupaus on luonut kiinnostusta sitä kohtaan liiketoiminnassa ja yhteiskunnan eri sektoreilla. Tiedon suurta määrää pidetään tärkeänä arvonluomisen mahdollistajana organisaatioille (Gandomi & Haider, 2015). Big datassa on alhainen arvotiheys (Zhu & Cai, 2015). Vain pieni osa alkuperäisestä tiedosta on siis arvokasta ja yksittäisen datayksikön arvo yksinään pieni, mutta suurta hyötyä saadaan analysoimalla suuria määriä tällaista tietoa (Gandomi & Haider, 2015). Arvolla big datan piirteenä on alun perin tarkoitettu taloudellista arvoa (Mikalef ym., 2018) ja se edustaakin usein liiketoiminnan näkökulmaa (Baesens ym., 2016), mutta sillä voidaan viitata myös sosiaaliseen arvoon, kuten hyvinvoinnin parantamiseen koulutuksen, terveydenhuollon ja turvallisuuden kautta (Günther ym., 2017; Jee & Kim, 2013). Tiedon todenmukaisuuden ja tiedosta saadun arvon välinen yhteys on suora ja selkeä, sillä laadultaan kelvoton tieto on huono lähtökohta arvon jalostamiselle (mm. Baesens ym., 2016; Ghasemaghaei & Calic, 2019; kuvio 1).



KUVIO 1 Big datan tärkeimmät ominaisuudet. Alhainen todenmukaisuus heikentää moninaisen, määrältään suuren ja nopeasti kertyvän big datan jalostamista arvoksi.

2.1.2 Tieto, informaatio, tietämys ja big data

Tieto on laaja käsite, jota voidaan jäsennellä eri tavoin (Laihonen ym., 2013). Usein käytetään Ackoffin (1989) hierarkiaan (data-information-knowledge-wisdom) perustuvaa jaottelua, jossa tiedon tasot ovat data, informaatio ja tietämys (Rowley, 2007). Termien käyttö on kuitenkin horjuvaa. Suomeksi sanoja data, tieto ja informaatio käytetään rinnakkain viittaamaan tietoon eri yhteyksissä. Englanniksi termien käyttö on erityisen epäloogista: tieto on ”data”, ”information” tai ”knowledge” riippuen kontekstista ja kirjoittajasta. (Laihonen ym., 2013; ks. esim. Wand & Wang, 1996; Batini ym., 2014). Tässä tutkielmassa käytetään yleisterminä näistä kaikista sanaa tieto. Data- ja informaatio-sanoja käytetään, kun on välttämätöntä erotella niin sanottu raakadata jalostetummista tiedon muodoista ja päinvastoin.

Ackoffin (1989) tiedon tasoista ylempi taso perustuu aina alemman tasoiselle tiedolle. Kun ”merkityksetöntä” ja runsasta dataa järjestellään ja muokataan siten, että se saa merkityksen, se muuttuu informaatioksi. Informaatio on rakenteistettua ja merkityksen saanutta dataa, joka on analysoitavissa. Tietämys perustuu dataan ja siitä jalostettuun informaatioon, ja se voi olla joko eksplisiittistä tai kokemukseen perustuvaa, hiljaista tietoa, jota ei ole tallennettuna tietojärjestelmissä. (Laihonen, 2013.)

Ackoffin (1989) jaottelu on keinotekoinen (Rowley, 2007; Jones, 2019). Kategoriat voivat olla pikemminkin jatkumo kuin toisistaan tiukasti eroteltavia tiedon tyyppejä (Rowley, 2007). Lisäksi tiedon lajien erittelyn perusteena on rakenteisuus ja merkityksellisyys eikä se sisällä esimerkiksi sellaisia ominaisuuksia kuin siirrettävyys (portability), sovellettavuus (applicability) ja toiminnan perusteeksi sopivuus (actionability) (Rowley, 2007). Näitä voitaisiin pitää tärkeinä ominaisuuksina big datan hyödyntämisessä. Big datan aikakaudella voidaan myös kyseenalaistaa kokemukseen perustuvan tietämyksen ylivertaisuus (ks. esim. McAfee & Brynjolfsson, 2012). Jaottelua voidaan kritisoida myös relativistisesta näkökulmasta käsin (ks. Jones, 2019).

Jos data määritellään positivistisesta näkökulmasta, todellisuuden kuvaajana, tiedolla ei teoriassa ole rajoja. Kitchinin (2014) määritelmän mukaan big data on laajuudeltaan tyhjentävää tietoa, joka pyrkii tavoittamaan kokonaisia populaatioita ja järjestelmiä mahdollisimman yksityiskohtaisella tasolla ja joka laajenee ja skaalautuu periaatteessa loputtomasti. Ei kuitenkaan ole olemassa ”merkityksetöntä” dataa. Datan jalostamiseksi informaatioksi ja tietämykseksi tarvitaan tietämystä, jonka avulla dataa voidaan kerätä, järjestellä ja jalostaa. Esimerkiksi tietoa käytännössä aina kerätään sen perusteella, mitä ymmärretään ilmiön luonteesta ja käytetään sen perusteella, mitä tietoa on saatavilla. Tietämys ajallisesti edeltää dataa ja on edellytys sen jalostamiselle ja taas uuden tietämyksen luomiselle. (Jones, 2019.) Tieto ei siis ole yksiselitteistä, staattista ja samaa aina ja kaikkialla, vaan relatiivista, suhteessa käyttökontekstiin ja -tilanteeseen.

Clarken (2016) mukaan big data on sekä dataa että informaatiota, jolloin datalla viitataan tiedonkeruun kontekstiin, informaatiolla kerätyn tiedon käyttökontekstiin ja datan relevanssiin siinä. Jones (2019) puolestaan erottelee ”tiedon periaatteessa” (data in principle), joka viittaa tallennettavaan tietoon ja ”tietoon käytännössä” (data in practice), tietoon, jota käytetään. Vaikka tietoa (big dataa) kertyy paljon, usein siitä vain pieni osa on käytettävissä. Tieto tulee olevaksi vain situationaalisten käytäntöjen, käsitteellistämisen, tallentamisen ja käytön, kautta. (Jones, 2019.) Big datan osalta juuri käytettävä tieto tai toisin sanoen informaatio ja se, miten se soveltuu käyttökontekstiinsa, on keskeinen (Clarke,

2016; Jones, 2019; De Mauro ym., 2016), sillä informaatio on big datan arvon lähde (De Mauro ym., 2016).

2.1.3 Big data -teknologiat

Big datan tallennus, hallinta, analysoiminen ja visualisointi vaativat uudenlaisia kehittyneitä teknologioita (Chen ym., 2012), sillä tiedon määrä ja rakenteettomuus tekee sen prosessoinnin perinteisillä teknologioilla vaikeaksi ja kalliiksi (Constantiou & Kallinikos, 2015; Davenport, Barth, & Bean, 2012; Gartner, 2012). Big data - teknologiat ovat kehittyneet aiempien ratkaisujen, kuten liiketoimintatiedon hallinnan (business intelligence), tiedonlouhinnan (data mining) ja tietovarastoinnin (data warehousing), pohjalta (Chen ym., 2012; Davenport, 2018).

Big datan varastointi edellyttää teknologioita, joiden avulla voidaan tallettaa suuria määriä vaihtelevan muotoista ja nopeasti kertyvää tietoa (Khine & Wang, 2018). Rakenteisten SQL-tietokantojen sijasta tietovarastoinnin standardina on skeematon NoSQL (Not-only-SQL) -lähestymistapa (Haseeb & Pattun, 2017). Prosessointi puolestaan vaatii runsaasti laskentatehoa. Pilvilaskentaa hyödyntämällä voidaan saada käyttöön kulloinkin tarvittava määrä laskentaresursseja, ja monet pilvipalveluyhtiöt tarjoavatkin analytics-as-a-service - eli AaaS-ratkaisuja (Demirkan & Delen, 2013). Riittävien laskenta- ja varastointiresurssien lisäksi big datan käsittely vaatii tehokkaita tietoverkkoja, jotka tukevat suurempia ja nopeampia tiedonsiirtoja (Xiong ym., 2013).

Big data -arkkitehtuurit pohjautuvat pääasiassa Apache Software Foundationin avoimen lähdekoodin projektille, Hadoopille (Highly Available Object Oriented Data Platform), joka perustuu Javaan (Hashem ym., 2015; Jee & Kim, 2013; Khine & Wang, 2018). Hadoopin tärkeimmät komponentit ovat tiedon varastoinnista vastaava HDFS (Hadoop Distributed File System)-levyjärjestelmä ja MapReduce-ohjelmointikehys, jotka liittyvät fyysisesti toisiinsa (Hashem ym., 2015). Ekosysteemiin kuuluu myös muita välineitä, kuten Hive, Hbase ja Mahout. Hadoop toimii sekä datan organisoijana että analytiikkavälineenä (Raghupathi & Raghupathi, 2014.) Se auttaa ratkaisemaan suurten aineistojen varastointiin, tietoon pääsyyn sekä yleiskustannusten hallintaan liittyviä ongelmia ja mahdollistaa hyvin nopean hajautetun prosessoinnin (Hashem ym., 2015; Jee & Kim, 2013; Khine & Wang, 2018), mutta on haasteellista asentaa, konfiguroida ja hallinnoida (Rahgupathi & Raghupathi, 2014). Hadoopia (MapReduce'a) tehokkaampi vaihtoehto big datan analysointiin on Apache Spark, klusterilaskennan väline, joka voi käyttää muun muassa Hadoopin levyjärjestelmää. (Lu, Hwang, & Huang, 2020.)

2.1.4 Tietoallasratkaisut

Perinteiset tietovarastot (data warehouse) eivät pysty vastaamaan big datan synnyttämiin haasteisiin. Niihin voidaan säilöä vain rajallinen määrä kaikesta organisaatioissa syntyvästä tiedosta. Kun tieto sijaitsee useissa erillisissä paikoissa, ns. siiloissa, siihen on vaikea päästä käsiksi, ja eri tietolähteitä on vaikea yhdistellä. Tämä on periaatteessa vanha ongelma, mutta tietoaltaat (data lakes) ratkaisevat sen tuomalla kaiken organisaatiossa syntyvän tiedon yhteen. (Khine & Wang, 2018.)

Tietoaltaalla ei ole vakiintunutta määritelmää tai arkkitehtuuria, mutta se määritellään yleensä varastoksi, jossa raakadataa säilytetään sen alkuperäisessä muodossaan (Ravat & Zhao, 2019). Tietoaltaan idea on, että kaikki organisaation tieto tallennetaan yhteen tietorakenteeseen eli tietoaltaaseen ilman monimutkaista prosessointia ja muokkaamista, joita tarvittaisiin tiedon lataamiseksi perinteiseen tietovarastoon. Tietoaltaaseen ladattava tieto voi olla rakenteista, rakenteetonta, heikosti rakenteista tai bittimuotoista ja eri tietoa voidaan ladata altaaseen eri aikataululla erissä, reaaliajassa tai tietovirtana. Tallennusvaiheessa tietoon yhdistetään metadataa. (Khine & Wang, 2018.)

Tietovarastoinnissa tietoa siirretään tietovaraston valmiiksi määriteltyihin SQL-tietokantarakenteisiin säännöllisissä erissä ETL-prosessin (extract-transform-load) kautta. Operationaalisten tietokantojen tieto siirretään (extract) ja prosessoidaan, puhdistetaan ja muokataan (transform) ennen sen lataamista (load) tietovarastoon. Varastoon ladattavan tiedon tulee siis olla rakenteista ja valmiiksi tietovarastoon sopivaksi muokattua. (Khine & Wang, 2018.) Tätä tietovarastoinnin ETL-prosessia tietoallasympäristössä vastaa ELT- tai EL-prosessi (Ravat & Zhao, 2019). Tiedon varastointi on edullista ja helppoa, koska tiedolta ei vaadita mitään ennalta määrättyä muotoa (Khine & Wang, 2018). Tiedon muokkaaminen on joustavaa ja tapahtuu käyttövaiheessa, ja myös tiedon ymmärtäminen jää käyttäjän tehtäväksi. Tämä vaatii sekä big data -teknologiaosaamista että sovellusalan tuntemusta. Tietoaltaasta voidaan yhdistellä hyödyntämätöntä, alkeellisella tasolla olevaa dataa rakenteiseen tietoon arvon luomiseksi. Epämääräinen rakenne ja avoimen lähdekoodin ratkaisut tekevät tietoaltaista kuitenkin haavoittuvia tietoturvan ja tiedon luottamuksellisuuden osalta, kun taas tietovarastot ovat varma ratkaisu tiedonhallinnan organisoinnin, tehokkuuden, tietoturvan ja tietoon pääsyn kontrollin kannalta, ja tieto on niissä semanttisesti yhtenäistä. (Khine & Wang, 2018.)

Tietoallas voi perustua erilaisille arkkitehtuureille. Yksinkertaisin niistä on litteä arkkitehtuuri, joka tallentaa kaiken raakadatan sen alkuperäisessä muodossa. Tämä arkkitehtuuri liittyy Hadoop-ympäristöön ja mahdollistaa runsaan ja heterogeenisen tiedon lataamisen alhaisin kustannuksin. Se ei kuitenkaan anna käyttäjien prosessoida tietoa eikä tallenna käyttäjien tekemiä operaatioita. Monimutkaisemmat arkkitehtuurit koostuvat useammasta pienemmästä tietoaltaasta (data ponds). (Ravat & Zhao, 2019.) Ravat'in ja Zhaon (2019) mukaan tietoallas ei korvaa tietovarastoja, sillä niillä on osittain erilaiset tavoitteet ja käyttäjät.

2.1.5 Big data -analytiikka ja päätöksenteko

Big data on arvotonta, jos siitä ei voida jalostaa mielekkäitä näkemyksiä päätöksenteon pohjaksi. Tähän organisaatiot tarvitsevat tehokkaita tiedonhallinnan ja analytiikan prosesseja tiedonhankinnasta aina sen tulkintaan saakka. Tiedonhallinta sisältää tiedon hankinnan, varastoinnin ja sen valmistelun ja hakemisen analyysiä varten (Gandomi & Haider, 2015). Analytiikka puolestaan tarkoittaa tiedon prosessoimista ja tekniikoita, joilla tietoa analysoidaan ja siitä saadaan jalostettua näkemyksiä sekä sosiaalista ja taloudellista arvoa (Gandomi & Haider, 2015; Günther ym., 2017). Wamba ym. (2015) määrittelevät big data -analytiikan big datan viiden ulottuvuuden (määrä, moninaisuus, nopeus, todenmukaisuus ja arvo) hallinnaksi, prosessoinniksi ja analysoinniksi, joiden tarkoituksena on luoda ideoita pysyvän arvon, yrityksen suorituskyvyn mittaamisen ja kilpailuetujen saavuttamiseksi.

Big datan myötä päätöksenteosta on tullut dynaamista (De Mauro ym., 2016). Tämä edellyttää organisaatioilta erityisesti analytiikan aseman uudelleen pohtimista ja samalla syvälle meneviä kulttuurisia muutoksia. Big dataa hyödyntävät yritykset sijoittavat analytiikan aiempaa kauemmas IT-funktiosta ja lähemmäs ydinliiketoimintaa sekä operationaalista toimintaa ja tuotantoa. Ne myös tarvitsevat perinteisestä analytiikasta olennaisesti poikkeavaa osaamista analysoidakseen jatkuvaa tiedon virtaa. (Davenport ym., 2012.) Big dataa hyödynnettäessä korkeatasoiseen IT- ja analytiikkaosaamiseen tulisi yhdistyä liiketoiminnan tuntemusta ja viestintätaitoja (Chen ym., 2012). Tärkeinä rooleina on mainittu datatieteilijä (Abbasi, Sarker, & Chiang, 2016; Davenport ym., 2012; Davenport, 2018), big data -insinööri ja tietoarkkitehti sekä IT- ja liiketoimintayksiköiden rajapinnassa toimiva yhteyshenkilö (Mikalef ym., 2018; Mikalef & Pateli, 2017).

Gandomi ja Haider (2015) käyvät kirjallisuuskatsauksessaan läpi big datan analyysitekniikoita, kuten tekstin, äänen, kuvan, videon ja sosiaalisen median tiedon analyysitekniikoita sekä ennustavaa analytiikkaa. Manyikan ym. (2011) ja Chenin ym. (2012) mukaan tavallisimmat menetelmät big datan prosessoinnissa ovat regressiomallit, klusterianalyysi, geneettiset algoritmit, signaalinkäsittely, luonnollisen kielen prosessointi, sosiaalisten verkostojen ja sentimenttien analyysi sekä tiedon visualisointi. Davenportin (2018) mukaan analytiikassa siirrytään kohti tekoälyn käyttöä, sulautettua ja automatisoitua analytiikkaa sekä kognitiivisia teknologioita, joiden hyödyntämisessä tarvitaan perinteisestä analytiikasta eroavia menetelmiä, kuten oppimisalgoritmeja. Syvällistä osaamista näiden tekniikoiden soveltamisen mahdollisuuksista ja rajoista ei juuri ole tällä hetkellä organisaatioiden saatavilla (De Mauro ym., 2016).

Holistinen tutkimus siitä, miten organisaatiot tuottavat arvoa big data-analytiikalla on niukkaa (Wang ym., 2019). Wamba ym. (2018), osoittivat resurssiperustaiseen näkemykseen pohjautuen, että big datan hallinta ja big data -infrastruktuuri sekä henkilöstön big data -osaaminen vaikuttavat yrityksen tulokseen, ja vaikutus välittyy osittain liiketoimintaprosessin dynaamisen kyvykkyyden kautta. Wangin ym. (2019) terveydenhuollon kontekstissa toteutetussa konfiguraatioteoriaan perustuvassa vertailevassa laadullisessa tutkimuksessa big data -analytiikkaa hyödyntävien organisaatioiden korkealaatuinen hoito liittyi erityisesti korkeatasoisiin analyyttisiin ja tiedon tulkitsemisen kyvykkyyksiin yhdistettynä tiedon integroinnin ja ennustamisen kyvykkyyteen sekä analytiikkahenkilöstön teknisiin taitoihin. Big data -analytiikalla oli suora parantava vaikutus organisaation toimintaan. (Wang ym., 2019.) Molempien tutkimusten mukaan organisaatiolta vaaditaan korkealaatuista teknologiaa, tiedonhallintaa ja osaamista, jos ne haluavat jalostaa big datasta arvoa.

Jotta tietoa voidaan käyttää tehokkaasti päätöksenteossa, koko henkilöstöä on kannustettava pitämään arvossa ja toteuttamaan huolellista tiedonhallintaa ja perustamaan päätöksensä tietoon (Buhl ym., 2013).Tämä edellyttää eritasoista dataosaamista kaikilla organisaation tasoilla (Wang ym., 2019). Hyödyntääkseen big dataa parhaalla mahdollisella tavalla organisaatioiden on siis käytävä läpi organisationaalisia ja kulttuurisia muutoksia, hankittava korkeatasoista analytiikan ja liiketoiminnan osaamista sekä uudistettava prosessejaan.

2.1.6 Tiedolla johtaminen

Big datan arvo ja merkitys on siinä, miten se parantaa organisaatioiden päätöksentekoa (Gandomi & Haider, 2015). Big datan myötä erityisesti julkisten palveluiden johtamisessa toivotaan voitavan siirtyä proaktiiviseen ja reaaliaikaiseen tiedon hyödyntämiseen, joka mahdollistaisi palvelujen paranevan tuottavuuden ja vaikuttavuuden (Jalonen, 2015). Puhutaan ”tiedolla johtamisesta”, joka käsitteenä on monitulkintainen (Jalonen, 2015). Usein sekä ”evidence based management” (tässä näyttöön perustuva johtaminen) että ”knowledge management” (tässä tietojohtaminen) kääntyvät suomeksi tiedolla johtamiseksi (ks. Jalonen, 2015; Hyppönen ym., 2012). Näiden käsitteiden juuret ovat kuitenkin erilaiset, ja niistä ensin mainittua käytetään erityisesti terveydenhuollon johtamisessa (ks. esim. Pfeffer & Sutton, 2006; Hyppönen ym., 2012). Molemmissa pyritään hankkimaan organisaation toiminnan ja päätösten perustaksi paras saatavilla oleva tieto.

Tietojohtamisen (knowledge management) juuret ovat resurssiperustaisessa ajattelussa (resource-based view), jossa tietoa pidetään yrityksen tärkeänä resurssina ja tiedon johtamista yrityksen menestystekijänä (Jalonen, 2015). Näyttöön perustuva johtaminen (evidence based management) on puolestaan kehittynyt näyttöön perustuvasta lääketieteestä (evidence based medicine), jolla on pitkät perinteet (Stewart, 2002). Näyttöön perustuvassa lääketieteessä lääkäri käyttää oman yksilöllisen kliinisen kokemuksensa lisäksi ulkopuolista tieteellistä näyttöä hoitopäätösten perustana (Sackett ym., 1996).

Englanninkielinen käsite ”knowledge management” on käännetty suomeksi tiedolla johtamiseksi, tietojohtamiseksi ja tietämyksenhallinnaksi (Jalonen, 2015). Riippuen koulukunnasta tai tutkimustraditiosta se on saanut erilaisia merkityksiä (ks. Laihonen ym., 2013). Suomessa kattokäsitteenä pidetään usein tietojohtamista. Sen lähtökohtana ovat organisaation toiminnassa kohdatut käytännön haasteet, ja siinä pyritään löytämään organisaation johtamiseen soveltuvia malleja ja työkaluja. Tietojohtamisen alle erotetaan toisistaan tiedon johtaminen ja tiedolla johtaminen. Ensin mainittu viittaa ”organisaation oppimiseen ja uusiutumiseen, uuden tiedon luontiin sekä tietovarantojen ja -virtojen hallintaan”. Toiseksi mainittu puolestaan ”toimintatapoihin, joilla organisaation tietoa jalostetaan ja hyödynnetään organisaation toiminnan johtamisessa”. (Laihonen ym., 2013.) Tiedolla johtamista voidaan siis pitää organisaation kohtaamien haasteiden ratkaisemisena tietoa keräämällä, jalostamalla ja hyödyntämällä.

Jalosen (2015) mukaan tietojohtaminen johtaa usein tiedon ylituotantoon. Ei riitä, että johtamisen perustana on paras käytettävissä oleva tieto, vaan sen tulee parantaa toimintaa vähentämällä tiedon puutteesta johtuvaa epävarmuutta tai paljosta tiedosta tai toiminnan monimutkaisuudesta johtuvaa monitulkintaisuutta. Ihanteena on relevantin ja epäolennaisen tiedon erottaminen toisistaan. (Jalonen, 2015.)

2.2 Tiedon laatu

Edellisessä alaluvussa käsiteltiin big dataa määrältään suurena, moninaisena ja nopeasti kertyvänä tietona, jolla on alhainen todenmukaisuus ja josta voidaan jalostaa runsaasti arvoa. Se määriteltiin tiedoksi, joka on sidoksissa keruu- ja käyttökonteksteihinsa, joissa myös siitä saatava hyöty tai arvo syntyy. Tiedon laatu vaikuttaa suuresti tiedon hyödyn-

tämiseen (Ghasemaghaei & Calic, 2019). Nyt siirrytäänkin käsittelemään tiedon laatua, sen määritelmiä ja tiedon laatu-ulottuvuuksia.

2.2.1 Tiedon laatu palveluna ja tuotteena

Laatu yleensä voidaan määritellä monesta eri näkökulmasta riippuen siitä, onko laadun kriteerinä jokin erinomaisuuden standardi, tuotteen tai palvelun arvo, vaatimusmäärittely vai asiakkaan odotukset (Nelson, Todd, & Wixom, 2005; Reeves & Bednar, 1994). Tiedon laatu voidaan eritellä datan ja informaation laatuun, jolloin kerätty ja tallennettu tieto on dataa ja käytettävä tieto informaatiota (Clarke, 2016).

Kahn, Strong ja Wang (1997) sekä myöhemmin Price & Shanks (2004) erottavat tuoteperustaisen ja palveluperustaisen näkemyksen tiedon laadusta. Tuoteperustainen tiedon laatu on datan laatua, sitä miten esimerkiksi tietokannoissa oleva data vastaa sitä ilmiötä, jota sen on tarkoitus ilmentää (laatu reaalimaailman vaatimusten noudattamisena). Tyypillisiä tiedon tuotelaadun kriteereitä ovat täydellisyys ja virheettömyys, joita mitataan objektiivisin mittarein. Pelkkä datatuote ei välttämättä vastaa tiedon käyttäjien tarpeisiin. Palvelusuuntautunut näkemys määrittelee informaation laadun sen perusteella, miten hyvin tietojärjestelmän tarjoama tieto vastaa käyttäjien tarpeisiin heidän itsensä määrittelemänä (laatu odotusten täyttämisenä). Tyypillisiä kriteereitä ovat ajantasaisuus, relevanssi ja saatavuus. Palveluperusteinen näkemys voi sisältää implisiittisesti tietotuotteen laadun. (Price & Shanks, 2004.)

Kun tieto tuotteena keskittyy datan ominaisuuksiin, tieto palveluna keskittyy siihen, mitä tiedolla tehdään. Se tuo mukaan tietoasiakkaan näkökulman. Tietoasiakkaat eivät erottele tiedon ja sitä jakavien laitteistojen ja sovellusten laatua toisistaan. Tietoa käytettäessä tiedon piilevät ominaisuudet, kuten sen helppokäyttöisyys, aggregoitavuus ja saatavuus, tulevat esiin. (Kahn ym., 2002.) Kun tutkitaan tiedon laatua palveluna, ei voida siis erotella ”puhdasta” tietoa, vaan on huomioitava kaikki se, mikä tiedon käyttäjän näkökulmasta vaikuttaa siihen, miten laadukasta tieto on. Tällaisia asioita voisivat olla esimerkiksi käyttäjän tiedot ja taidot, teknologiaympäristö ja tiedon käyttötarkoitus.

Tiedon laadun määritelmästä ei ole konsensusta ja määritelmiä on monia (Batini ym., 2014). Määritelmiin vaikuttaa se, lähestytäänkö sitä empiirisestä, käytännön asiantuntijan vai teoreettisesta näkökulmasta. Teoreettisesti johdetut laatumääritelmät ja kriteerit ovat täsmällisempiä ja sisäisesti koherentimpia, mutta eivät tarpeeksi laajoja, sillä ne eivät ota huomioon kuluttajan näkökulmaa, vaan vain tiedon tuoteaspektit. (Price & Shanks, 2004.) Yleisintä on määritellä tiedon laatu empiirisestä näkökulmasta sopivuudeksi käyttöön (fitness-for-use). Tällaiset määritelmät ovat kontekstisidonnaisia eikä laatuominaisuuksille, kuten tiedon virheettömyys tai täydellisyys ole ontologisesti yhtenäistä määritelmää. (Liaw ym., 2013.)

Tiedon laatu on keskeinen tietojärjestelmien omaksumista ja käyttöä selittävä tekijä tietojärjestelmätieteessä (Nelson ym., 2005). Se sisältyy teoreettisena käsitteenä esimerkiksi DeLonen ja McLeanin (1992, 2003) malliin, jolla he selittävät tietojärjestelmän myönteisten tai kielteisten (netto)vaikutusten syntymistä. Mallissa tietojärjestelmän vaikutuksia ei rajata vain järjestelmän suoraan käyttäjään, vaan menestymisen mittana on organisaation ja jopa yhteiskunnan saama hyöty (DeLone & McLean, 1992, 2003). DeLonen ja McLeanin

(2003) määritelmä on lähellä laatua arvona: tieto on sitä parempilaatuista, mitä hyödyllisempää se on tai mitä enemmän siitä voidaan jalostaa arvoa (vrt. Reeves & Bednar, 1994).

2.2.2 Tiedon laatu-ulottuvuudet ja -ominaisuudet

Tiedon laatu-ulottuvuuksien tutkimus on yksi monista tiedon laadun tutkimuksen haaroista ja erityisesti tietojärjestelmätieteen ja laadullisen tutkimuksen alaa (Rao, Gudivada, & Raghavan, 2015; Sadiq, Yeganeh, & Indulska, 2011). Vaikka tiedon laadun määritelmänä sopivuus käyttötarkoitukseen on laajasti hyväksytty, tiedon laatuominaisuuksien rakenne ja nimeämiskäytäntö vaihtelee suuresti. Usein tiedon laatu eritellään erilaisiin laatuominaisuuksiin tai hyvän laadun piirteisiin, kuten virheettömyyteen, täydellisyyteen ja ajantasaisuuteen. Määritelmät voivat myös alakohtaisesti poiketa toisistaan. (Laranjeiro, Soydemir, & Bernardino, 2015.)

Kirjallisuudessa tiedon laadusta uraauurtavana voidaan pitää Wangin ja Strongin (1996) tutkimusta, jossa empiirisesti, tietoasiakkaiden mielipiteiden pohjalta muodostettiin luokittelu tiedon laatuominaisuuksista ja niiden muodostamista laatu-ulottuvuuksista (Laranjeiro ym., 2015; Wang & Strong, 1996). Tutkimuksessa löydettiin yli sadan tiedon laatuominaisuuden joukosta 15 tärkeintä laatuominaisuutta faktorianalyysin keinoin (Wang & Strong, 1996). Nämä luokiteltiin tiedon laadun neljään ”laatu-ulottuvuuteen”: sisäiseen laatuun (intrinsic data quality), saatavuuden laatuun (accessibility data quality), kontekstuaaliseen laatuun (contextual data quality) ja representationaaliseen laatuun (representational data quality) (ks. Taulukko 1). Mallista jätettiin pois muutamia ominaisuuksia, jotka vastaajat sijoittivat epäjohdonmukaisesti eri ulottuvuuksille ja joita ei pidetty kovin tärkeinä (Wang & Strong, 1996). Nämä laatuominaisuudet on lisätty taulukon 1 loppuun. Wangin ja Strongin (1996) tutkimus korosti tekniikan asiantuntijanäkökulman sijaan juuri tiedon käyttäjän näkökulmaa: tekniset vaatimusmäärittelyt täyttävä, korkealaatuinen, analytiikkatyökaluilla helposti käsiteltävä tieto, ei välttämättä ole laadukasta (Wang & Strong, 1996; Clarke, 2016).

Tiedon laatua ei siis voi mitata tarkastelemalla vain sen sisäisiä laatuominaisuuksia, kuten virheettömyyttä, irrallaan tiedon käyttötarkoituksesta, käyttäjistä ja käyttöprosessista. Se on monimutkaisten organisationaalisten prosessien tulosta ja aina sidoksissa tiedon käyttökontekstiin, jossa hyödyllisyys ja käytettävyys ovat tärkeitä laatuaspekteja. (Strong ym., 1997.) Tästä näkökulmasta on myös ymmärrettävää, että tiedon sisäinen laatu-ulottuvuus sisältää ominaisuuksia, joita ei teknisestä näkökulmasta ajatella sellaisiksi. Tiedon *sisäinen laatu* sisältää datan objektiiviset ominaisuudet (Taleb, Serhani, & Dssouli, 2018). Käyttäjän näkökulmasta sisäinen laatu on kuitenkin laajempi asia, sillä tiedon lähteen uskottavuus ja maine liittyvät läheisesti tiedon virheettömyyteen ja objektiivisuuteen, vaikka niitä ei voidakaan suoraan havaita tarkastelemalla tietoa (Wang & Strong, 1996).

Tiedon *kontekstuaalinen laatu* puolestaan määräytyy sen perusteella, miten hyvin tietoa voidaan hyödyntää sen käyttökontekstissa (Strong ym., 1997). Tiedon käyttökonteksti voi muuttua, kun asiakkaan työtehtävissä tapahtuu muutoksia tai samaa tietoa käytetään erilaisiin tarpeisiin. Tällöin tapahtuu muutoksia myös tiedon laatuominaisuuksissa. (Strong ym., 1997.) Jos sisäinen laatu on lähellä datan laatua, kontekstuaalisesta laadusta voidaan puhua informaation laatuna (Taleb ym., 2018).

TAULUKKO 1 Tiedon laatu-ulottuvuudet (Wang & Strong, 1996).

Sisäinen laatu	
Virheettömyys	Missä määrin tieto on oikeaa, luotettavaa ja todistettavasti virheetöntä
Objektiivisuus	Missä määrin tieto on ennakkoluulotonta ja puolueetonta
Uskottavuus	Miten uskottavaa tieto on
Maine	Missä määrin tietoon luotetaan tai sitä suuresti arvostetaan sen lähteen tai sisällön perusteella
Saatavuuden laatu	
Saatavuus	Missä määrin tieto on saatavilla tai helposti ja nopeasti haettavissa
Pääsyn tietoturva	Missä määrin pääsyä tietoon voidaan rajoittaa ja säilyttää sen tietoturva
Kontekstuaalinen laatu	
Relevanssi	Missä määrin tieto on sovellettavissa ja avuksi käsillä olevassa tehtävässä
Lisäarvo	Missä määrin tieto on hyödyllistä ja sen käytöstä on etua
Ajantasaisuus	Missä määrin tiedon ikä on sopiva käsillä olevaan tehtävään
Täydellisyys	Missä määrin tieto on sopivan kattavaa, syvällistä ja laajaa käsillä olevaan tehtävään
Tiedon määrä	Missä määrin tiedon määrä on sopiva
Representationaalinen laatu	
Tulkittavuus	Missä määrin tieto on sopivaa kieltä ja yksiköt ja tiedon määrittelyt ovat selkeitä
Ymmärtämisen helppous (ymmärrettävyys)	Missä määrin tieto on selkeää eikä siinä ole monitulkintaisuutta ja se on helposti ymmärrettävissä
Esittämisen tiiviys	Missä määrin tieto on tiiviisti esitetty ilman että se on liiallista (lyhyt mutta täydellinen ja täsmällinen esitystapa)
Esittämisen yhdenmukaisuus (yhdenmukaisuus)	Missä määrin tieto esitetään aina samassa formaatissa ja on yhteensopivaa aiemman tiedon kanssa
Mallista pois jätetyt laatuominaisuudet	
Jäljitettävyys	Missä määrin tieto on hyvin dokumentoitua, tarkistettavissa ja sen lähde on helposti todennettavissa
Tiedon ja tietolähteiden moninaisuus	Missä määrin tietoa on saatavissa useista erilaisista tietolähteistä
Kustannustehokkuus[1]	Missä määrin sopivan tiedon keräämisen hinta on kohtuullinen
Teknisen käsittelyn helppous	Missä määrin tieto on helposti hallittavissa ja manipuloitavissa (päivitettävissä, siirrettävissä, aggregoitavissa, sovitettavissa käyttötarkoitukseen, jäljennettävissä)
Joustavuus	Missä määrin tieto on laajennettavissa, mukautettavissa ja helposti sovellettavissa muihin tarkoituksiin

[1] Tämän tutkimuksen empiirisessä osassa kustannustehokkuus on määritelty sekä tiedonkeruun että tiedon käsittelyn kustannustehokkuudeksi.

Tietoasiakkaan näkökulmasta tiedon *saatavuus* on laaja käsite, se sisältää helppouden, jolla he pystyvät manipuloimaan tietoa tarpeisiinsa. Esimerkiksi lääketieteellinen blob-muodossa tallennettu kuvantamistieto ei ole saatavilla lääkärin näkökulmasta, jos hänen jos hän ei pysty analysoimaan sitä käytössään olevin ratkaisuin. Aineisto, joka koostuu eri tietolähteiden tiedoista, voi olla teknisesti saatavilla, mutta käyttäjät eivät koe niin, koska samanlaiset tiedot on määritelty, mitattu ja esitetty eri tavoin. Koodattu lääketieteellinen tieto on teknisesti saatavilla tekstinä, mutta tietoasiakkaiden mielestä se ei ole saatavilla, koska he eivät osaa tulkita koodeja. Suuri määrä tietoa on teknisesti saatavilla, mutta asiakkaat eivät koe niin, koska siihen käsiksi pääsemiseen menee runsaasti aikaa. (Strong ym., 1997.)

Representationaalinen tiedon laatu sisältää formaattiin ja merkitykseen liittyviä aspekteja. Tieto tulisi esittää tiiviisti ja johdonmukaisesti, mutta myös niin, että sen tulkitseminen ja ymmärtäminen on helppoa. Esimerkiksi tietokannassa formaatti liittyy syntaksiin ja merkitys semanttiseen yhteensopivuuteen. (Wang & Strong, 1996.) Esimerkkinä voisi käyttää valuutan merkitsemistä euroina tai dollareina.

Wangin ja Strongin (1996) empiirinen luokittelu on kohdannut kritiikkiä. Laatukriteerien johtamista käyttäjien palautteesta ja luokittelemista kategorioihin voidaan pitää epäjohdonmukaisena, epäselvänä ja tarkoitushakuisena (Price & Shanks, 2004). Esimerkiksi Eppler (2001) on kritisoinut Wangin ja Strongin (1996) luokittelua siitä, että laatukriteerien välisiä riippuvuuksia (esim. uskottavuus edellyttää mainetta) ei ole perusteltu. Riippuvuuksien tuominen näkyviin auttaisi ymmärtämään, miten tiedon laatuongelmat vaikuttavat toisiin laatukriteereihin (Warwick ym., 2015). Mallin kaikki kriteerit eivät myöskään ole yleisiä ja sovellettavissa kaikille aloille (esim. objektiivisuus) (Price & Shanks, 2004). Koska sekä teoreettinen että empiirinen näkökulma on yksinään riittämätön laadun arvioimisen kannalta, osa tutkijoista on pyrkinyt luomaan synteesejä, joissa molempia näkökulmia huomioidaan (esim. Price & Shanks, 2004).

Laranjeiron ja Soudemirin (2015) kirjallisuuskatsauksen mukaan yleisimmin kirjallisuudessa käytetyt tiedon laatuominaisuudet ovat saatavuus (accessibility), virheettömyys (accuracy), täydellisyys (completeness), yhdenmukaisuus (consistency) ja ajantasaisuus (currency). ISO/IEC 25012 -standardi (ISO 2008) määrittelee edellä mainituista neljä viimeisintä ulottuvuutta yhdessä uskottavuuden (credibility) kanssa tietotuotteen sisäisiksi laatuominaisuuksiksi. Tämän lisäksi standardissa on vielä tiedon sekä sisäisistä laatuominaisuuksista että järjestelmästä riippuvia laatuominaisuuksia, jotka ovat saavutettavuus (accessibility), sääntöjenmukaisuus, luottamuksellisuus, tehokkuus, tarkkuus, jäljitettävyys ja ymmärrettävyys, sekä yksinomaan järjestelmästä riippuvat laatuominaisuudet saatavuus (availability), siirrettävyys ja palautuvuus. (ISO 2008.) Batini ym. (2014) kritisoivat standardia siitä, että se ottaa huomioon vain rakenteisen tiedon, joten sen avulla ei voi arvioida skeematonta tietoa, kuten tekstidokumentteja.

Seuraavaksi siirrymme käsittelemään tiedon laatuongelmia ja sitä, miten ne syntyvät tiedon käyttökonteksteissa.

2.3 Tiedon laatuongelmat ja niiden syyt

Tiedon laatuongelmista puhutaan usein etenkin tietovarastoinnin yhteydessä huonona (bad) tai likaisena (dirty) tietona, joka on puhdistettava (cleanse, clean, scrub) tiedon laadun varmistamiseksi (esim. Laranjeiro ym., 2015; Rahm & Do, 2000). Keskittyminen varastoidun tiedon sisäisiin laatuongelmiin ei kuitenkaan riitä, sillä laatu liittyy monimutkaisiin organisaationaalisiin konteksteihin (Strong ym., 1997). Onkin riski, että tiedon laatuongelmat rajataan ainoastaan tiedon sisäisiksi ongelmiksi, jolloin organisaation kriittisimmät tiedon laatuongelmat voivat jäädä huomiotta (Clarke, 2016). Tiedon organisaationaaliseen käyttökontekstiin liittyvät ja siitä nousevat laatuongelmat nostivat esille Strong ym. (1997). Heidän mukaansa tiedon laatuongelma on mikä tahansa, missä tahansa laatu-ulottuvuudessa kohdattu ongelma, joka tekee tiedosta kokonaan tai suurelta osin käyttöön sopimattoman. Tällainen voi olla esimerkiksi puuttuva tieto (Strong ym., 1997).

Rahmin & Don (2000) tutkimus on uraauurtava erityisesti heterogeenisten tietovarastoaineistojen yhdistämiseen liittyvistä laatuongelmista (Laranjeiro & Soydemir, 2015). He jakavat tiedon sisäiset laatuongelmat neljään luokkaan sen perusteella, johtuvatko ne yhdestä vai useammasta tiedon lähteestä ja ovatko ne instanssi- vai skeemakohtaisia (Taulukko 2). Heidän määrittelemänsä yhdestä tiedon lähteestä juontuvat laatuongelmat vaikuttavat laatuominaisuuksista yleisimmin tiedon saatavuuteen ja virheettömyyteen, useampien tietolähteiden käytöstä johtuvat tiedon laatuongelmat puolestaan saatavuuteen ja yhdenmukaisuuteen. (ks. Laranjeiro ym., 2015; Rahm & Do, 2000.)

Laatuongelmat vaikuttavat yleensä yhtä aikaa heikentävästi moniin tiedon laatu-ulottuvuuksiin (Laranjeiro ym., 2015; vrt. Strong ym., 1997). Esimerkiksi puuttuva tieto on Rahmin ja Don (2000) mukaan yhdestä lähteestä johtuva instanssikohtainen laatuongelma, joka vaikuttaa sekä tiedon täydellisyyteen että virheettömyyteen[2] (taulukko 2). Rahmin ja Don luokittelu auttaa analysoimaan tiedon laatua sen prosessoinnin näkökulmasta, mutta voi antaa laatuongelmista staattisen kuvan.

Tiedon kontekstin huomioiminen laatuongelmien synnyssä tuo esille sen, millaisten syy-seuraussuhteiden tulosta tiedon laatu ja sen heikkoudet ovat, miten esimerkiksi organisaation tekniset resurssit, ihmisten uskomukset, käyttäjien tietotarpeet ja tiedon keruun ja käsittelyn prosessit synnyttävät tiedon laatuongelmia, joista puolestaan voi aiheutua muita tiedon laatuongelmia. Strongin ym. (1997) empiirinen tutkimus tiedon kontekstuaalisesta laadusta perustuu Wangin ja Strongin (1996) määrittelemiin tiedon laatu-ulottuvuuksiin. Heillä esimerkiksi edellä mainittu tiedon puute voi syntyä käyttäjien muuttuvista tietotarpeista tai prosessoinnin virheistä ja puolestaan vaikuttaa aineiston relevanssiin tiedon käyttökontekstissa (Strong ym., 1997). Tiedon puute voi siis olla sekä relevantin aineiston puuttumista kokonaan että aineistosta puuttuvia tietoalkioiden arvoja, vaikka perinteisesti siihen viitataan lähinnä jälkimmäisessä merkityksessä puuttuvana tietona.

Seuraavaksi tarkastellaan tiedon laatuongelmia Wangin ja Strongin (1996) viitekehyksessä. Ensimmäinen laatu-ulottuvuus on *tiedon sisäinen laatu*, joka kattaa paitsi virheettömyyden myös tiedon uskottavuuden ja objektiivisuuden (Wang & Strong, 1996). Tiedon

[2] Taulukko 2 on yhdenmukainen Laranjeiron ym. (2015) kanssa. Rahmin ja Don (2000) artikkelin perusteella voisi kuitenkin päätellä, että puuttuva tieto voi olla myös saatavuusongelma.

TAULUKKO 2 Tiedon laatuongelmien luokittelu Rahmin ja Don (2000) ja Laranjeiron ym. (2015) mukaan.

Ongelmatyypit		**Tiedon laatuongelmat**	**Saatavuus**	**Virheettömyys**	**Täydellisyys**	**Yhdenmukaisuus**	**Ajantasaisuus**
Lähde	**Taso**						
Yksittäinen	**Instanssi**	Puuttuva tieto		x	x		
		Virheellinen tieto		x			
		Kirjoitusvirheet		x			
		Epäselvä tieto	x	x			
		Asiaankuulumaton tieto	x			x	
		Vanhentunut aikatieto		x			x
		Arvot väärissä kentissä	x	x	x	x	
		Virheelliset viitteet		x			
		Duplikaatit	x				
	Skeema	Toimialueen rikkomus		x			
		Funktionaalisen riippuvuuden rikkomus		x			
		Väärä tietotyyppi	x			x	
		Viite-eheyden rikkomus	x	x	x	x	
		Kaksoisarvojen eston rikkomus		x			
Monta	**Instanssi**	Rakenteelliset konfliktit	x			x	
		Erilaiset sanajärjestykset	x			x	
		Erilaiset aggregaatiotasot	x	x		x	
		Ajallinen yhteensopimattomuus		x		x	x
		Erilaiset yksiköt	x			x	
		Erilaiset esittämistavat	x			x	
	Skeema	Synonyymien käyttö	x				
		Homonyymien käyttö	x				
		Erikoismerkkien käyttö	x				
		Erilaiset koodittamisformaatit	x			x	

sisäiseen laatuun vaikuttaa Strongin ym. (1997) tutkimuksen mukaan se, että samaa tietoa saadaan useista, toisistaan poikkeavista lähteistä, ja se, että tiedontuotantoon liittyy subjektiivista harkintaa. Tietojen uskottavuus voi heiketä, kun tietoa aineistojen eroavaisuuksien syistä kertyy. Kun sitten tiedon käytössä on ongelmia, käsitys siitä, että tiedon sisäinen laatu on huono, muuttuu yleiseksi tiedoksi, huonoksi maineeksi. Tiedosta ei saada juuri lisäarvoa eikä sitä käytetä. (Strong ym., 1997.) Rahmin ja Don (2000) luokitteluun sisältyy Wangin ja Strongin (1996) mallista virheettömyys. Heidän mukaansa vihreettömyyteen vaikuttavat useimmat vain yhdestä tiedon lähteestä johtuvat tiedon sisällön ja rakenteen laatuongelmat (Taulukko 2). Tällaisia ovat muun muassa puuttuva tai virheellinen tieto tai tietokannan kaksoisarvojen eston rikkominen (uniqueness constraint violation). Kun useita tiedon lähteitä yhdistetään, virheettömyys voi heiketä aineistojen toisistaan poikkeavien aggregaatiotasojen ja ajallisen yhteensopimattomuuden vuoksi. (Laranjeiro ym., 2015). Useiden tiedonlähteiden käyttäminen siis tuo esille tiedonkeruun ja yksittäisten tiedonlähteiden ongelmia, joiden vuoksi tiedon maine kärsii. Monien lähteiden käyt-

täminen myös vaikuttaa heikentävästi aineiston virheettömyyteen, mikäli tiedot koskevat eri ajankohtaa tai tietoalkioiden luokittelussa on eroja.

Wangin ja Strongin (1996) laatu-ulottuvuuksista *saatavuus* sisältää tiedon saatavuuden ja pääsyn tietoturvan. Kummankin heikkoudet voivat hidastaa, estää tai hankaloittaa tiedon käyttöä (Wang & Strong, 1996). Rahmin ja Don (2000) luokittelussa lähes kaikki monesta lähteestä johtuvat ongelmat, kuten synonyymien käyttö, heikentävät tiedon saatavuutta, mutta myös yhdestä lähteestä johtuvilla ongelmilla, kuten epäselvällä tiedolla, tuplilla tai väärällä tietotyypillä, on tiedon saatavuutta heikentävä vaikutus. Kuitenkin kaikki monesta lähteestä johtuvat yhdenmukaisuuden ongelmat ovat samalla saatavuuden ongelmia. (Rahm & Do, 2000.) Erilaisten aineistojen yhdistäminen usein estää, hankaloittaa tai hidastaa tiedon käyttöä.

Strongin ym. (1997) tutkimuksen mukaan saatavuusongelmia aiheutui järjestelmien alhaisista laskentaresursseista, jotka voivat estää pääsyä tietoihin, ja tietoturvasäädöksistä, joiden vuoksi tiedon käyttöönsä saadakseen on käytettävä aikaa ja vaivaa. Saatavuusongelmia aiheutui myös, jos aineisto sisälsi useiden erikoisalojen tietoja ja oli koodattua, jolloin sen ymmärrettävyys ja tulkittavuus oli huono, eikä käyttäjä pystynyt itsenäisesti käyttämään tietoa. Lisäksi suuren tietomäärän prosessointi voi olla hidasta ja viedä liikaa aikaa, ja näin saatavuus kärsii, kun tieto ei ole ajanmukaisesti käytettävissä. Käyttäjä ei siis eri syistä pääse tietoon käsiksi, ei pysty tulkitsemaan ja ymmärtämään sitä tai saa sitä käyttöön riittävän nopeasti. (Strong ym., 1997.)

Sekä Strongin ym. (1997) että Rahmin ja Don (2000) perusteella tiedon esittämistapaan, tiedon *representationaaliseen laatuun*, liittyvä saatavuuden heikkous on keskeinen tiedon laatuongelma. Representationaalinen laatu sisältää Wangin ja Strongin (1996) mallissa tulkittavuuden, ymmärtämisen helppouden sekä esittämistavan yhdenmukaisuuden ja tiiviyden. Rahmilla ja Dolla (2000) tätä ulottuvuutta vastaa tiedon yhdenmukaisuus (Rahm & Do, 2000). Heikko representationaalinen laatu heikentää, kuten edellä tuli ilmi, saatavuuden laatua, eli esimerkiksi sitä, miten helposti ja nopeasti käyttäjä voi ymmärtää tiedon sisällön. Se heikentää myös *kontekstuaalista laatua*, johon Wang ja Strong (1996) sisällyttävät lisäarvon, relevanssin, ajankohtaisuuden, täydellisyyden ja riittävän tiedon määrän. Rahmilla ja Dolla (2000) tätä ulottuvuutta vastaavia ominaisuuksia ovat ajankohtaisuus ja täydellisyys.

Strong ym. (1997) tunnistavat erityisesti tiedon epätäydellisyyden ja tiedon epäyhtenäisen esittämistavan, heikon relevanssin ja alhaisen lisäarvon vaikutuksen tiedon huonoon kontekstuaaliseen laatuun. Puuttuva tai epätäydellinen tieto, joka voi johtua tiedon prosessoinnissa tapahtuvista virheistä tai tietoasiakkaiden muuttuvista tarpeista, heikentää tiedon relevanssia. Eri tiedonlähteiden epäyhtenäinen tiedon esittämistapa puolestaan vaikuttaa siihen, että tiedon aggregointi ja yhdistely on hankalaa ja tiedosta saadaan vain vähän lisäarvoa. Näin vähäinen lisäarvo ja heikko relevanssi alentavat tiedon kontekstuaalista laatua. (Strong ym., 1997.) Relevanssi ja lisäarvo ovat siis välillisiä laatumääreitä, jotka ovat riippuvaisia tiedon muista laatumääreistä käyttökontekstissa. Rahmin ja Don (2000) mukaan vanhentunut ajallinen tieto tai eri aineistojen ajallinen eroavaisuus heikentävät ajantasaisuutta. Puuttuva tieto puolestaan vaikuttaa tiedon täydellisyyteen. Nämä tiedon laatuongelmat vaikuttavat heillä samalla heikentävästi tiedon virheettömyyteen (Laranjeiro & Soydemir, 2015.) Ei-ajantasainen tai kontekstissaan puutteellinen tieto on siis samalla virheellistä tietoa, kun puuttuva tieto määritellään puuttuviksi tietoalkioiden arvoiksi.

Strongin ym. (1997) tutkimus osoittaa, miten tiedon laatu ja laatuongelmat ovat yhteydessä toisiinsa ja syy-seuraussuhteessa tiedon käyttökontekstiin. Rahmin ja Don (2000) tiedon laatuongelmien luokittelu täydentää ja täsmentää lisäksi ymmärrystä tiedon sisäisten ja teknisten laatuongelmien vaikutuksista tiedon laatuun.

2.4 Big datan laatu ja laatuongelmat

Big data, erityisesti sen alhainen todenmukaisuus ja moninaisuus, on monimutkaistanut ja pahentanut tiedon laatuongelmia (Rao ym., 2015). Raon ym. (2015) mukaan tietovirrat, erilaiset tietotyypit ja monet tiedon toimittajat, koneen generoima rakenteeton tieto ja yhdistämisongelmat ovat joko erityisiä big data -kontekstille tai niiden haasteellisuus kasvaa siinä. Big dataa hyödynnettäessä sen laatuun tuleekin kiinnittää erityistä huomiota (Baesens ym., 2016; Ghasemaghaei & Calic, 2019), koska tiedon huono laatu on yrityksille taloudellisesti erittäin kallista (Cichy & Rass, 2019), ja huonolaatuisen big datan perusteella tehtyjen päätösten seuraukset voivat esimerkiksi terveydenhuollossa olla inhimillisesti ja yhteiskunnallisesti ajatellen kohtalokkaita (Clarke, 2016.)

Tiedon laadun merkitys korostuu big datan aikakaudella. Ennen 2010-luvun puoliväliä erityinen big datan laatua koskeva tutkimus on kuitenkin lähes puuttunut (Laranjeiro & Soydemir, 2015; Baesens ym., 2016), mutta sitä on alkanut ilmestyä aivan viime vuosina. Kirjallisuus painottuu tiedon tekniseen laadunarviointiin ja -hallintaan analytiikan näkökulmasta (Cappiello ym., 2018; Merino ym., 2016; Taleb ym., 2018; Talha, El Kalam, & Elmarzouqi, 2019). Keskeistä on kustannustehokas ja järjestelmien suorituskyvyn huomioonottava arvon jalostaminen tiedosta (Merino ym., 2016; Taleb ym., 2018; Ardagna, ym., 2018; Surbakti ym., 2020). Kiinnostus on siirtynyt tiedon prosessoinnin alkupäästä sen loppupäähän. Kun perinteinen teknisen tiedon laadun tutkimus korostaa datasyötteiden valvomista ja kontrollointia, big datan laatua koskeva tutkimus keskittyy tiedon hyödyntämisen tuloksiin (Loshin, 2014, Surbaktin ym. (2020) mukaan). Tiedon käytön konteksti korostuu, sillä vain siinä tiedon arvo voidaan määritellä (Merino ym., 2016).

Varhaisimpia big datan laadun viitekehyksiä on Merinon ym. (2016) malli (Laranjeiro & Soydemir, 2015). Se perustuu ISO/IEC 25012- ja 25024- standardeihin ja korostaa big datan laatuominaisuutena tiedon sopivuutta (adequacy) analyysin ja analyytikon tarpeisiin arvonluomisen näkökulmasta (Quality-in-Use). Big datalla on Merinon ym. (2016) mukaan kolme sopivuuspiirrettä tai -ulottuvuutta: kontekstuaalinen sopivuus (missä määrin eri aineistoja voi käyttää samalla sovellusalueella), ajallinen sopivuus (tiedon tuottamisen, analysoinnin ja ymmärtämisen aikahaarukan yhteneväisyys) ja operationaalinen sopivuus (missä määrin aineistoa on teknisesti mahdollista analysoida käytössä olevan big data -ratkaisun avulla vaikuttavasti ja tehokkaasti). Tavoitteena on ”riittävän hyvälaatuinen” tieto. (Merino ym., 2016.)

Big datan laatu-ulottuvuuksien ja -ominaisuuksien määrittelyyn voidaan ehdottaa valmiita perinteisen tiedon laadun viitekehyksiä (esim. Taleb ym., 2018; Talha ym., 2019; Ghasemaghaei & Calic, 2019) tai muokata niitä big datalle sopiviksi (esim. Merino ym., 2016; Clarke, 2016). Tiedon käyttökontekstin painottuessa laatumallien kehittely on kuitenkin paljolti alakohtaista (Rao ym., 2015). Ardagnan ym. (2018) mukaan big datan tietotyyppien, lähteiden ja sovellusten tulisi vaikuttaa sen laatuominaisuuksiin ja laadunarviointiin. Esimerkiksi suuri lähteiden määrä tekee luottamuksesta ja uskottavuudesta tärkei-

tä laatuominaisuuksia. Täydellisyys pitäisi tietovirtojen kontekstissa ymmärtää kahdesta eri perspektiivistä: tietyn lukeman täydellisyytenä ja kokonaisten tietovirtojen täydellisyytenä. (Ardagna ym., 2018.) Talhan ym. (2019) mukaan eheys tarvitsisi selkeän määritelmän tiedon laatunäkökulmasta.

Big data -ympäristöt ovat lisänneet tiedon laadun hallinnan haasteita (Merino ym., 2016). Big datan prosessoinnin ratkaisut poikkeavat tietovarastointiin pohjautuvista järjestelmistä ja voivat itsessään edesauttaa varastoidun ja käytettävän tiedon laatuongelmien syntyä. Big datan varastointi tietoaltaissa alkeellisessa muodossa, jota muokataan vasta sitä käyttöä varten valmisteltaessa, on omiaan aiheuttamaan tietokatkoksia, varsinkin kun big data on tyypillisesti lähtöisin useilta eri tiedontoimittajilta ja sitä käsittelevät erilaiset välikädet (Rao ym., 2015). Tieto myös käy läpi useita muunnoksia ja "mutkittelee" eri sovellusten kautta, jolloin laatuvirheet voivat monistua ja kasautua (Rao ym., 2015). Yleinen ongelma on metadatan puute, eli tiedot esimerkiksi siitä, mikä oli muuttujan skaala tiedonkeruussa tai tietojen alkuperä, puuttuvat. Tärkeää metadataa olisivat myös tiedot tiedon laatuun vaikuttavista tekijöistä, kuten, mitä merkitysmuutoksia ajan myötä on tapahtunut (Clarke, 2016.)

Clarkella (2016) metadatan ja tiedon laatua varmistavien prosessien ja menettelytapojen puute itsessään ovat informaation laatuongelmia, jotka koskevat käytettävää big dataa. Myös Rao ym. (2015) korostavat tiedon alkuperän, alkuperäisten lähteiden ja tiedolle tehtyjen muunnosten kirjaamista. Buhlin ym. (2013) mukaan ilman tiedon laadunhallinnan prosesseja ja vastuukysymysten ratkaisemista edistys teknologisessa infrastruktuurissa, analyyttisissä työkaluissa ja liiketoimintamalleissa ovat tietoon perustuvan päätöksenteon kannalta arvottomia. Big datalle kehitetyt tiedonlaadunhallintamallit ovat kuitenkin vasta alhaisella kypsyystasolla eivätkä big data -arkkitehtuurit tue laadunhallinnan prosesseja (Taleb ym., 2018). Raon ym. (2015) mukaan suurin ongelma on se, että automatiikkaa big datan laatuongelmien ratkaisemiseen ei ole, joten analysoiminen nojaa merkittävässä määrin datan manuaaliseen puhdistamiseen. Ongelmia voi aiheutua myös tietoturvan ja tiedon laadunhallinnan ristiriitaisista vaatimuksista (Talha ym., 2019).

Empiirinen tutkimus big datan laadun vaikutuksesta päätöksenteolle arvokkaan informaation jalostamiseen on tähän mennessä ollut vähäistä (Ghasemaghaei & Calic, 2019). Ghasemaghaei ja Calic (2019) selvittivät survey-tutkimuksessaan pohjoisamerikkalaisten yritysten data-analyytikkojen ja it-managerien käsityksiä. Tutkimuksen tulosten mukaan big dataa käyttävien yritysten päätöksenteon laatu on hyvää eli tuottaa taloudellista hyötyä vain siinä määrin, kuin tiedon laatu on hyvä. Vain hyvälaatuisesta big datasta voidaan jalostaa arvokasta informaatiota (ns. diagnostisesti korkeatasoista informaatiota), ja arvokas informaatio on puolestaan edellytyksenä hyvälle päätöksenteolle. (Ghesemaghaei & Calic, 2019.)

Tiedon sisäisellä, kontekstuaalisella ja representationaalisella laadulla oli Ghasemaghaein ja Calicin (2019) tutkimuksessa merkitsevä vaikutus arvokkaan informaation jalostamiseen. Näistä sisäisen laadun vaikutus oli suurin, ja vain siihen big datan käyttö vaikutti heikentävästi. Tiedon laatua on vaikea varmistaa, kun suuria määriä tietoja yhdistellään nopeasti useista lähteistä. (Ghasemaghei & Calic, 2019.) Big datan laadun arvioiminen, puhdistaminen ja muokkaaminen järkevässä ajassa onkin haasteellista nykyisillä tiedon prosessoinnin välineillä (Cai & Zhu, 2015). Koska sisäisen laadun vaikutus datasta jalostetun informaation arvoon oli suurin ja merkitsevin, Ghasemaghei ja Calic (2019) suosittelevat kiinnittämään siihen eniten huomiota big dataa hyödynnettäessä, esimerkiksi

puhdistamalla, suodattamalla ja yhdistelemällä tietoa jo tiedonkeruun varhaisessa vaiheessa. Yritysten on, heidän mukaansa, halutessaan luoda big datasta arvoa, järkevämpää kiinnittää huomiota varastoidun tiedon laatuun kuin varastoida suuria määriä toisiinsa liittymätöntä dataa. (Ghasemaghei & Calic, 2019.)

Vaikka big datan laadun tarkka varmistaminen joka vaiheessa on teoriassa ainoa keino varmistaa sen perusteella tehtyjen johtopäätösten oikeellisuus, pragmaattisesta näkökulmasta se ei ole mahdollista. Tiedon laadun hallinta perinteisin menetelmin ei sovi big datalle, koska big data -ympäristöissä tiedon laadun heikkouksien etsimiseen ja korjaamiseen tarvittavien sääntöjen määrä olisi valtava, eikä löydettyjen ongelmien korjaaminen olisi toteutettavissa tai edes mahdollista (Merino ym., 2016; Ardagna ym., 2018). Big dataa käytettäessä on siis välttämättä tehtävä hallittuja kompromisseja runsaan, nopean ja moninaisen tiedon ja sen laadun välillä. Baesensin ym. (1996) mukaan big dataa käyttävien organisaatioiden tulisikin määritellä, mikä on riittävä virhemarginaali tiedoille ja varmistaa tiedon laatu kaikilla tiedon prosessoinnin ja analyysin tasoilla alkaen alkuperäisen tiedon keräämisestä, tiedon varastoinnista ja tiedon hakemisesta sekä sen valmistelemisesta analyysiä varten.

Hyötyäkseen big datasta organisaatioiden on kiinnitettävä huomiota big datan ja siitä jalostetun tiedon laatuun, sillä ilman riittävän korkeaa laatua jalostaminen ei ole taloudellisesti kannattavaa eikä paranna päätöksentekoa. Määrältään suuren, nopean ja moninaisen tiedon laadun arviointi ja varmistaminen on mahdotonta samaan tapaan kuin perinteisen tiedon perinteisessä tietovarastoinnissa, joten laadunvarmistuksen prosesseilla ja menettelytavoilla sekä hyvällä metadatalla on erityisen suuri merkitys jalostettaessa tietoa big data -ympäristöissä.

3 TIEDON MONINAISUUS KÄRJISTÄÄ POTILASTIEDON LAATUONGELMIA TOISSIJAISESSA KÄYTÖSSÄ

Terveydenhuollon tietona voidaan pitää kaikkea terveydenhuollon käytettävissä olevaa, potilaiden tai populaatioiden terveyteen tai hoitoon liittyvää tietoa (Auffray ym., 2016; Mehta & Pandit, 2018). Sen keskeisin tietotyyppi on *kliininen tieto* (esim. Kruse ym., 2016), jolla tarkoitetaan hoitohenkilökunnan potilaan hoidon yhteydessä kirjaamia ja tallentamia *sairauskertomustietoja* potilaan terveydestä, hoidosta ja toimenpiteistä sekä sairaaloiden ja klinikoiden hallinnollista ja taloustietoa (ks. Szlezák ym., 2014).

Terveydenhuollon kliinistä tietoa ja big dataa käytetään usein toissijaisiin tarkoituksiin eli muuhun kuin siihen tarkoitukseen, johon se on alun perin kerätty (Clarke, 2016). Tässä luvussa tarkastellaan kliinisen tiedon ja - big datan laatuongelmia ja niiden syitä käytettäessä tietoa tutkimukseen ja päätöksentekoon. Aluksi käydään läpi terveydenhuollon big datan määritelmiä, lähteitä ja tiedon tyyppejä sekä tiedon toissijaista käyttöä. Luvun loppupuolella käsitellään terveydenhuollon tiedon laatukriteereitä sekä laatuongelmia ja niiden syitä tiedon toissijaisessa käytössä. Koska potilastietojen laatuongelmat ovat terveydenhuollon big datan keskeinen laatuongelmien lähde (ks. Hoffman, 2014), luvun loppuosassa käsitellään potilastietojen laatuongelmia ja analysoidaan niiden syitä toissijaisessa käytössä, määrittelivätpä artikkelin kirjoittajat tiedon big dataksi tai perinteiseksi tiedoksi (vrt. Richesson, Horvath, & Rusincovitch, 2014).

Kirjallisuushaku on tehty käyttäen Google Scholar -, PubMed-, Scopus- ja Web of Science -tietokantoja tavoitteena löytää mahdollisimman relevanttia ja tuoretta kirjallisuutta. Hakulauseissa yhdisteltiin alla olevia hakusanoja sekä käytettiin lumipallotekniikkaa:

Tiedon laatu ja laatuongelmat: *data quality problems, data quality issues, data quality*

Sovellusalue ja tiedon tyyppi: *healthcare, electronic medical record, EMR, electronic health record, EHR, clinical data, patient record, medical record, clinical record, clinical information*

Toissijainen käyttö ja big data: *secondary use, reuse, big data*

3.1 Big datan määritelmä ja tyypit terveydenhuollossa

Lääketieteellisessä kirjallisuudessa, erityisesti genomiikan erikoisalalla, termin big data käyttö on yleistynyt vuodesta 2011 (Baro ym., 2015). Terveydenhuollon big datan operationaalisesta määritelmästä ei ole kirjallisuudessa yhteisymmärrystä, mutta eri määritelmillä on yhteisiä elementtejä (Mehta & Pandit, 2018). Ne perustuvat Mehtan ja Panditin (2018) kirjallisuuskatsauksen mukaan tiedon ominaisuuksiin, sen sisältämiin tietotyyppeihin, analytiikan ja tiedonhallinnan työkaluihin tai siihen, miten hyvin tieto sopii analysoitavaksi. Terveydenhuollon big data on Mehtan ja Panditin (2018) mukaan määrältään suurta, moninaista, nopeaa, sen todenmukaisuus on alhainen, ja siitä voidaan jalostaa arvoa (Mehta & Pandit, 2018).

Terveydenhuollon aineistot voidaan luokitella big dataksi esimerkiksi tiedon määrän tai sen merkityksen perusteella. Baron ym. (2015) mukaan big data on ennen kaikkea määrältään suurta riippumatta siitä hyödynnetäänkö tietoa keruukontekstissaan vai toissijaisiin tarkoituksiin. Sitä vastoin Jonesin (2019) mukaan terveydenhuollossa, erityisesti genomiikassa, on aineistoja, jotka voidaan määritellä big dataksi pelkästään tiedon kasvaneen merkityksen perusteella, vaikka ne olisivat teknisin termein perinteistä tietoa. Liyanagen ym. (2014) mukaan terveysaineiston koko on aikasidonnainen määre ja riippuvainen teknologisesta kehityksestä, eikä sinänsä tee aineistosta big dataa. Toisaalta suuret aineistot eivät ole lääketieteessä uutta, mutta aineistojen analyysimenetelmät ovat kehittyneet (Szlezák ym., 2014.)

Dinovin (2016) määritelmän mukaan terveydenhuollon big datalla on kaksi tärkeää piirrettä: energia ja elinkaari. Energia tarkoittaa yhdistelmäaineiston kokonaisvaltaista informaatiosisältöä, joka on suurempi kuin yksittäisen aineiston ja sitä hyödyllisempi analysoinnin näkökulmasta (Dinov, 2016). Elinkaarella Dinov (2016) puolestaan viittaa siihen, että terveydenhuollon big datan arvo heikkenee eksponentiaalisesti ajassa. Vaikka Dinov (2016) ei nojaa tavallisimpiin big datan ominaisuuksiin, voidaan ajatella, että energia on johdettavissa big datan moninaisuuden ja arvon käsitteistä. Elinkaari sen sijaan viittaisi tiedon kiihtyvään vanhenemisnopeuteen. Myös sen voi katsoa viittaavan arvoon, sillä terveyden big datasta analysoimalla saatava hyöty vähenee ajan myötä. Dinovin (2016) määritelmä siis korostaa aineistojen yhdistämisen ja tuoreuden merkitystä, jotta niiden analysoimisesta olisi mahdollisimman paljon hyötyä.

Terveydenhuollon big datan arvot ja käyttötarkoitus ovat erilaisia kuin liiketoiminnan, sillä terveydenhuollossa pyritään nopean kilpailuedun sijaan löytämään kestäviä ratkaisuja (Jee & Kim, 2013). Jeen ja Kimin (2013) mukaan terveydenhuollon big dataa määrittelevät nopeuden, määrän ja moninaisuuden sijaan siiloutuneisuus (silo), tietoturva (security) ja moninaisuus. Nämä ominaisuudet heijastelevat big datan käytön haasteita terveydenhuollossa. Terveydenhuollon tiedolle on ominaista, että se sijaitsee siiloutuneena sairaaloiden ja viranomaisten omissa rekistereissä ja tietovarastoissa. (Jee & Kim, 2013.)

Osa terveydenhuollon big datan määritelmistä liittyy tiiviisti analytiikkaan ja arvon luomiseen (Mehta & Pandit, 2018). Bates ym. (2014) määrittelevät big datan suureksi määräksi moninaista tietoa, jolla on potentiaalia kertyä nopeasti, sekä analytiikaksi, joka on yhteyksien löytämistä tiedosta sekä niiden kommunikoimista. Galetsin ym. (2020) mukaan big data on liian monimuotoista ja hankalaa käsitellä perinteisin laitteistoin ja ohjelmistoin, ja sen analysoimiseksi tarvitaan big data -analytiikkaa. Tämä tarkoittaa useimmiten tiedon

prosessointia ja analysointia Hadoop-ympäristössä (Galetsi ym., 2020; Raghupathi & Raghupathi, 2014b).

Terveydenhuollossa tuotetaan monenlaista digitaalista tietoa. Auffray ym. (2016) määrittelevät "terveyden big datan" suureksi määräksi hyvin monimuotoista biologista, kliinistä sekä ympäristötekijöihin ja elintapoihin liittyvää tietoa, jota on kerätty aina yksilöistä suuriin kohortteihin saakka suhteessa heidän terveys- ja kuntotasoonsa yhtenä tai useampana ajankohtana." Big datan lähteet terveydenhuollossa ovat hyvin heterogeenisia (taulukko 3). Terveydenhuollon big data sisältää sekä sairaaloiden ja klinikoiden sisäistä että niiden ulkoista tietoa. Ulkoisia lähteitä ovat esimerkiksi lääketieteelliset julkaisut, diagnostiikka- ja lääkeyritykset sekä julkishallinto. Potilaiden itse tuottamaa tietoa saadaan käyttöön verkossa toimivien terveysportaalien, sosiaalisen median sivustojen ja lääketieteellisten laitteiden kautta. (Mehta & Pandit, 2018.)

Terveydenhuollon kliininen tieto (ks. taulukko 3) ja erityisesti sähköinen sairauskertomus (electronic health record, EHR) on keskeistä tietoa terveydenhuollon big data -analytiikassa ja kliinisessä tutkimuksessa (Galetsi ym., 2020; Richesson, Hovarth, & Rusincovitch, 2014; Kruse ym., 2016). Sähköinen sairauskertomus syntyy eri tahojen potilasasiakirjoista (electronic medical record, EMR) ja on potilaan elinaikana kerättyä yksityiskohtaista sähköistä tietoa hänen terveydestään ja hoidostaan (Bonimi, 2016; Szlezák ym., 2014; Kruse ym., 2016). Potilasasiakirjojen ja sairauskertomuksen määrittely on vaihtelevaa. Suomessa potilasasiakirjoihin luetaan kaikki potilaan hoitoon liittyvät laaditut tai saapuneet asiakirjat ja tekniset tallenteet, kuten röntgenkuvat (Lääkäriliitto, 2020). Kirjallisuudessa ei usein erotella sähköistä sairauskertomusta ja potilasasiakirjoja, vaan molemmista puhutaan sähköisenä sairauskertomuksena (esim. Jetley & Zhang, 2019; Weiskopf & Weng, 2013). Suomessa sairauskertomuksesta käytetään myös nimeä potilaskertomus (ks. Lääkäriliitto, 2020).

3.2 Terveydenhuollon tiedon toissijainen käyttö

Terveydenhuollon tiedon ensisijaiset käyttötarkoitukset ovat tarkoituksia, joihin tieto on alun perin kerätty. Näitä ovat potilaiden hoito ja laskutus. (Botsis ym., 2010.) Toissijainen käyttö on tiedon käyttöä muuhun kuin ensisijaiseen tarkoitukseen, kuten biolääketieteelliseen tutkimukseen, julkisen terveydenhuollon kehittämiseen ja vakuutustoimintaan (Hoffman, 2014; Hoffman & Podgurski, 2013). Arvoa saadaan tiedon keräämisestä yhteen suuriksi aineistoiksi ja moninaisen tiedon yhdistämisellä. Esimerkiksi lääketieteellisessä tutkimuksessa ja julkisen terveydenhuollon kehittämisessä keskeisessä roolissa ovat potilas- ja hoitotietoa keräävät rekisterit (esimerkiksi syöpärekisterit), sillä niihin on koottuna paljon ja kattavasti tietoa yksittäisistä potilaista (Auffray ym., 2016). Tehokkaita hoitoja eri sairauksiin voidaan kehittää terveystiedon yhdistämisellä genomiikan ja muiden "omiikoiden" tietoon (Costa, 2014). Tällaisen toissijaisiin tarkoituksiin käytettävän big datan avulla halutaan tuottaa vaikuttavampaa ja kustannustehokkaampaa hoitoa ja luoda ratkaisuja julkisen talouden rahoitusvajeeseen (Feldman & Martin, 2013).

TAULUKKO 3 Terveydenhuollon tiedonlähteet (Taulukon lähtökohtana Szlezák ym., 2014; Mehta & Pandit, 2018 mukailemana. Lisäyksiin on merkitty lähdeviite.)

Tyyppi	**Laji**	**Kuvaus**	**Lähde**
Kliininen tieto	Sähköiset potilas-asiakirjat, sairauskertomukset[1]	Yksityiskohtainen potilaaseen liittyvä tieto (lääkärin määräykset, lääkitykset, hoitohistoria, diagnoosit[1] ja oirekuvaukset[1])	Sairaalat ja klinikat
	Diagnostiikka	Diagnostiset tulokset (kuvantaminen ja laboratoriomittaukset)	Laboratoriot Radiologian osastot
	Omiikka	Molekyylitieto (genomiikka, proteomiikka, transkriptomiikka ja metaboliikka)	Diagnostiikkayritykset
	Lisätieto	Hallinnollinen tieto (sairaalaan otto, sairaalasta lähtö, siirto) ja taloustieto (korvausvaatimukset)	Sairaalat ja klinikat Tiedon aggregoijat
Korvausvaatimukset	Lääketieteelliset vaatimukset	Lääketieteellisiin vaatimuksiin liittyvä tieto (toimenpiteet, sairaalapäivät, vakuutussopimusten yksityiskohdat)	Maksajat Tiedon aggregoijat
	Lääkemääräyksiin liittyvät vaatimukset	Lääkekorvaukset (lääke, annostus, kesto)	Maksajat Tiedon aggregoijat
Kliininen tutkimus	Kliiniset kokeet	Koeasetelma (kohdejoukko, koko, vastemuuttujat)	Lääkeyritykset Lääketieteen julkaisut
Potilaiden tuottama tieto	Sosiaalinen media	Yhteisöjen keskustelut	Verkon terveysportaalit Sosiaalisen median sivustot
	Henkilökohtaiset sensorit	Hyvinvointi- ja elintapatieto (älypuhelimet, kuntoseurannan laitteet)	Laitetiedon järjestelmät
	Kyselytieto	Klinikoiden potilailtaan keräämä tieto hoidon tuloksista (PROM- eli patient reported outcome measures -kyselyt[2]	Klinikat[2]

[1] Cirillo & Valencia, (2019), [2] Ehrenstein ym., (2017)

Big data -ilmiö on lisännyt kiinnostusta kliinisten potilastietojen käyttöön erityisesti lääketieteessä. Kliininen tutkimus tukee lääkitysten ja laitteiden, diagnostisten välineiden ja lääketieteellisten hoitojen turvallisuutta ja tehokkuutta. Sen sisällä ovat yleistyneet ei-kokeelliset, havaintotutkimusasetelmat, joissa hyödynnetään eri instituutioiden kliinisen ja hallinnollisen tiedon varastojen ja sairauskertomusjärjestelmien tietoa. (Richesson, Horvath, & Rusincovitch, 2014.) Sairauskertomustiedon etuna tutkimuskäytössä on, että suuria aineistoja voidaan käyttää ilman perinteisen kliinisen tutkimuksen vaatimia kustannuksia ja tehottomuutta. Retrospektiiviseen tutkimukseen ei tarvita potilaiden rekry-

tointia tai erillistä tiedonkeruuta, ja tarjolla on näkymä todellisen potilaspopulaation monimuotoisuuteen. Tieto ei kuitenkaan välttämättä sovellu toissijaiseen tarkoitukseen eikä ole laadultaan tutkimukseen riittävän korkeatasoista. (Weiskopf & Weng, 2013.) Koska sähköinen sairauskertomus on terveydenhuollon keskeisintä toissijaisiin tarkoituksiin käytettävää tietoa ja terveydenhuollon big datan ”ydintietoa” (Kruse ym., 2016; Galetsi ym., 2020), sen laadulla on suuri merkitys big datan hyödyntämisessä (Kruse ym., 2016).

3.3 Terveydenhuollon tiedon ja sairauskertomustiedon laatukriteerit toissijaisessa käytössä

Sairauskertomustieto poikkeaa tieteellisistä aineistoista monin tavoin. Siinä on usein monimutkaista pitkittäistietoa, päällekkäisyyksiä ja laajasti narratiivista tekstiä. (Weiskopf ym., 2013b.) Se ei siis ole konventionaalista rakenteista tietoa. Esimerkiksi laatuongelmiksi useimmiten luokiteltavat yhdenmukaisuuden puute ja tuplarivien esiintyminen voivat olla itsessään merkityksellisiä: ne voivat kertoa esimerkiksi toisistaan poikkeavista lääketieteellisistä mielipiteistä tai väärin tallennetuista tai useista koetuloksista. Lisäksi pelkästään tallennetun tiedon laatuvaatimukset eivät pysty varmistamaan tiedon sopivuutta tiettyyn käyttötarkoitukseen. (Orfanidis, Bamidis & Eaglestone, 2004.)

Tiedon laadunarvioinnilla tarkoitetaan usein tutkimusaineistojen laadunarviointia analyysiä varten. Weiskopfin ja Wengin (2013) systemaattisen kirjallisuuskatsauksen mukaan sairauskertomustiedon laadunarvioinnissa käytettyjen laatukriteerien määritelmät ovat vaihtelevia ja osin päällekkäisiä. Yleisimmin arvioidut laatuomaisuudet olivat *täydellisyys* ja *virheettömyys*. Yhdessä *ajantasaisuuden* kanssa ne olivat katsauksen perusteella sairauskertomustiedon perustavanlaatuiset laatuominaisuudet toissijaisessa käytössä. Lisäksi *konkordanssi* (järjestelmän sisäisten elementtien yhtäpitävyys toistensa tai ulkoisen lähteen kanssa) ja *uskottavuus* (yhdenmukaisuus yleisen lääketieteellisen tiedon kanssa) olivat osa tiedon laadunvarmistusta täydellisyyden, virheettömyyden ja ajantasaisuuden edustajina, jos niitä ei voitu arvioida suoraan. (Weiskopf & Weng, 2013; Weiskopf ym., 2017.)

Tärkein sairauskertomustiedon laatuominaisuus Weiskopfin ja Wengin (2013) mukaan on täydellisyys. Weiskopf ym. (2013a) jakavat sairauskertomustiedon täydellisyyden neljään tyyppiin, joista vain dokumentaation täydellisyys on tiedon sisäinen laatuominaisuus ja muut täydellisyyden määritelmät eli laajuuden, tiheyden ja ennustava täydellisyys riippuvat käyttökontekstista (taulukko 4).

TAULUKKO 4 Sähköisen sairauskertomuksen täydellisyyden määritelmät (Weiskopf ym., 2013b).

Täydellisyyden tyyppi	**Määritelmä**
Dokumentaation täydellisyys	Sairauskertomus sisältää kaikki potilaasta tehdyt havainnot
Laajuuden täydellisyys	Sairauskertomus sisältää kaikki tietyn tyyppiset tiedot
Tiheyden täydellisyys	Tietojen, esimerkiksi mittauskertojen, esiintymistiheys tai -määrä sairauskertomuksessa on riittävän suuri
Ennustava täydellisyys	Sairauskertomus sisältää riittävät tiedot ilmiön ennustamiseksi

Weiskopfin ja Wengin (2013) tutkimuksen perusteella voidaan päätellä, että sähköisen sairauskertomustiedon laadunarvioinnissa keskitytään useimmiten tietojen tarkistamiseen ja

validointiin melko kapeasta näkökulmasta. Ne keskittyivät Wangin ja Strongin (1996) määrittelemistä laatu-ulottuvuuksista osaan tiedon sisäistä (virheettömyys, uskottavuus) ja kontekstuaalista (täydellisyys, ajantasaisuus) laatua (Weiskopf & Weng, 2013). Lisäksi konkordanssi, jota Weiskopf ja Weng (2013) eivät lue mihinkään Wangin ja Strongin (1996) laatu-ulottuvuuteen kuuluvaksi, on lähellä representationaaliseen laatu-ulottuvuuteen kuuluvaa yhdenmukaisuutta.

Tiedon laadunarvioinnin menetelmiä sairauskertomusaineistojen toissijaisessa tutkimuskäytössä olivat Weiskopfin ja Wengin (2013) mukaan toisesta lähteestä peräisin olevan aineiston käyttäminen laadun kultaisena standardina, aineiston sisäisten dataelementtien vertailu, dataelementin olemassaolon mittaaminen, tietolähteiden yhtäpitävyys, aggregoidun tiedon jakaumien tai tunnuslukujen vertailu, tiedon validiteetin arviointi eri tekniikoin ja kirjausten lokitietojen läpikäynti. Suurimmassa osassa aineiston tutkimuksia nojattiin kultaiseen standardiin, intuitiiviseen ymmärrykseen tiedon laadusta ja ad hoc -tyyppisiin tiedon laadun arvioinnin menetelmiin (Weiskopf & Weng, 2013). Kultaisen standardin käyttö on ongelmallista, koska eri tarkoituksiin koostetuille, suurille ja varsinkin tunnistetiedoista puhdistetuille aineistoille on vaikea löytää sopivia vertailuaineistoja (Weiskopf & Weng, 2013).

Terveydenhuollon tiedon laadunarviointiin kehitetyt viitekehykset ja menetelmät ovat hyvinkin vaihtelevia. Laadunarvioinnin välineitä terveydenhuollon tiedolle ovat kehittäneet muun muassa Warwick ym. (2015) ja Bai ym. (2018), sähköisen sairauskertomuksen toissijaiseen käyttöön Johnson ym., (2015) ja Kahn ym., (2016) ja tutkimuskäyttöön Weiskopf ym. (2013b) ja Weiskopf ym. (2017). Laatuominaisuuksien määritelmät eri malleissa poikkeavat yleensä ontologisesti toisistaan, koska niillä mitataan sopivuutta käyttöön eri tarkoituksissa (Liaw ym., 2013).

Liawin ym. (2013) mukaan tulisi pyrkiä teoreettisesti loogisiin ja toisensa poissulkeviin käsitteisiin ja yleispätevään tiedon laatukriteeristöön. Esimerkiksi Weiskopfin ym. (2017) mittaristoon on valittu Weiskopfin ja Wengin (2013) kirjallisuuskatsauksen laatuominaisuudet, ja sen on tarkoitus toimia yleispätevästi erilaisten sairauskertomusaineistojen arvioinnissa. Warwick ym. (2015) puolestaan pyrkivät laadun arvioimiseen laajasta, intuitiivisia, teoreettisia ja empiirisiä kriteereitä yhdistelevästä näkökulmasta. Heidän laatukriteereitään olivat saatavuus, relevanssi, virheettömyys, luotettavuus, ajantasaisuus, selkeys, vertailtavuus, koherenssi, validiteetti ja luottamuksellisuus (Warwick ym, 2015). Laatuominaisuuksilla on riippuvuussuhteita, ja ne on heidän mukaansa tärkeää eksplisiittisesti eritellä. (Warwick ym., 2015, vrt. Price & Shanks, 2004.)

Bai ym. (2018) muodostivat tutkimuksensa tapausorganisaation liiketoimintatiedon hallinnan vaatimusten perusteella tiedon laatukriteeristön, joka perustui aiemman kirjallisuuden pohjalta tehdylle klusteroinnille. Klusterit olivat paikkansapitävyys (accuracy), täydellisyys (completeness), saavutettavuus (accessibility), yhdenmukaisuus (concistency), ei-päällekkäisyys (non-redundancy), luettavuus (readability, käyttökelpoisuus (usefulness) ja luottamus (trust). Huomionarvoista on, että ajantasaisuus kuuluu paikkansapitävyyden klusteriin ja relevanssi täydellisyyden klusteriin. Klustereissa on myös useita Wangin ja Strongin (1996) mallille rinnakkaisia ominaisuuksia, mutta esimerkiksi käyttökelpoisuus ja relevanssi on eroteltu tarkemmin toisistaan. Laadun arviointi perustuu mallissa sekä kvantitatiiviselle mittaamiselle että käyttäjien subjektiivisille arvioille.

Terveydenhuollon ja sairauskertomustiedon laatua mitataan siis hyvin erilaisin ja eri tavoin määritellyin mittarein. On pyrkimyksiä luoda yleispäteviä laatumittareita ja toisaal-

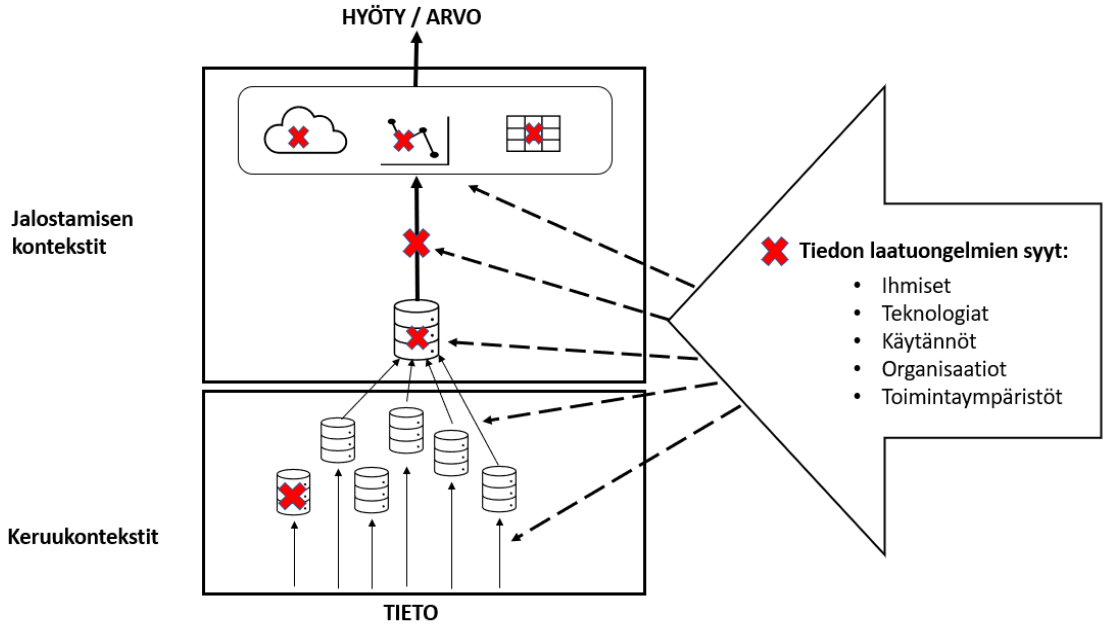
ta vastata laadunarvioinnin tarpeisiin tietyssä ympäristössä. Sopivuus käyttöön on subjektiivista, eikä sitä voida havainnoida suoraan tietoa tarkastelemalla. Joskus tiedon laadun arvioinnin kriteerinä pidetään varsinaisten tiedon laatukriteerien ohella tiettyjen menettelytapojen käyttöä, jolloin on kiistanalaista, onko tällöin kuitenkaan kyse itse tiedon laatuominaisuuksista (Weiskopf & Weng, 2013). Sairauskertomuksen laatukriteerinä voisi käyttää potilaan ja lääkärin vuorovaikutuksen laatua (Porter & Mandl, 1999), joka on epäilemättä edellytys laadukkaan tiedon keräämiselle ja voi olla syynä tiedon laatuongelmiin, mutta sen perusteella ei suoraan voida päätellä, miten laadukasta järjestelmiin tallennettu tieto on. Clarken (2016) viitekehyksessä informaation laatutekijät sisältävät ohjauksen (controls) eli riittävät liiketoiminnan prosessit varmistamassa sen, että datan ja informaation laatutekijät otetaan huomioon ennen datan käyttöä. Tässäkin on kyse tiedon laadun edellytyksistä, joiden puute voi aiheuttaa tiedon laatuongelmia. Jos ajatellaan koko tiedon elinkaarta aina analyysin tulosten tulkintaan asti, samantyyppisiä laadun edellytyksiä voisivat olla hyvän tieteellisen käytännön noudattaminen ja pätevien tutkimusasetelmien ja –menetelmien valinta (ks. Hoffman & Podgurski, 2013).

3.4 Kliinisen tiedon ja big datan laatuongelmat ja niiden syyt toissijaisessa käytössä

Terveydenhuollon tiedon laatuongelmat voivat syntyä missä tahansa tiedon elinkaaren vaiheessa, karkeasti jaoteltuna joko tietoja tallennettaessa tai tietoa käytettäessä (Clarke, 2016; Sukumar ym., 2015, kuvio 2). Ne voivat olla datatason laatuongelmia, joita voidaan arvioida tietokannan loogista rakennetta tai aineistoja tarkastelemalla (Laine ym., 2015). Tiedon laadunvarmistus tapahtuukin usein vain tällä tasolla (Orfanidis ym., 2004). Osa ongelmista liittyy kuitenkin tiedon tallentamisen tai käytön konteksteihin (Laine ym., 2015; Clarke, 2016). Käytännössä tiedon laatuongelmien syyt ovat samalla big datan laatuongelmien syitä (ks. esim. Hoffman, 2014), mutta ongelmat voivat olla big datassa vaikeampia, sillä tiedon moninaisuus kärjistää epätodenmukaisuudesta aiheutuvia ongelmia (Molinari & Nollo, 2020).

Sairauskertomustiedon laatuongelmia terveydenhuollon tiedon toissijaisessa tutkimuskäytössä ovat tutkineet muun muassa Bayley ym. (2013), Hersh ym. (2013) ja Hoffman ja Podgurski (2013). Bayley ym. (2013) ja Hersh ym. (2013) tutkivat sähköisen potilaskertomuksen laatua vertailevassa hoidon vaikuttavuustutkimuksessa. Bayleyn ym. (2013) mukaan haasteita olivat puuttuva ja virheellinen tieto, tieto, jota ei pystytä tulkitsemaan, palveluntarjoajien ja eri ajankohtien välinen tiedon epäyhdenmukaisuus sekä koodaamaton vapaa teksti. Hersh ym. (2013) pitivät ongelmina tiedon virheitä, epätäydellisyyttä, erilaisissa muunnoksissa heikentynyttä tiedon merkitystä, sitä, että vapaa teksti ei ole palautettavissa tutkimusta varten, monista lähteistä johtuvaa tiedon tuntematonta alkuperää, riittämätöntä tarkkuustasoa ja sitä, että tietoa ei ole kerätty tutkimusprotokollien mukaisella tavalla.

Hoffman ja Podgurski (2013) tutkivat sairauskertomus- ja geneettistä tietoa sisältävien biolääketieteellisten tietokantojen tutkimuskäytön haasteita. He pitivät tärkeimpinä



KUVIO 2 Tiedon keruu- ja jalostamisprosessi terveydenhuollon tiedon toissijaisessa käytössä. Laatuongelmia syntyy kaikissa vaiheissa, ne ovat moninaisia ja johtuvat monista syistä.

tiedon huomattavia virheitä, epätäydellisyyttä ja hajanaisuutta, tiedon ja tieteellisen päättelyn systemaattisia harhoja sekä tarkoitushakuisten ja harhaanjohtavien tutkimustulosten riskiä (Hoffman & Podgurski, 2013). Potilaskertomustietoa sisältävään aineistoon liittyy vaikeitakin tiedon laadun haasteita, jotka heijastuvat analyysin tuloksiin, päättelyyn ja tutkimuksen vaikutuksiin saakka, ja joita voidaan parantaa tai vaikeuttaa tiedon jalostamisvaiheessa.

Kirjallisuudessa big datan laatuongelmia on käsitelty pääosin terveydenhuollon kliinisen ja erityisesti sairauskertomustiedon osalta. Seuraavassa käydään läpi terveydenhuollon tiedon laatuongelmien syitä toissijaisessa käytössä jaoteltuna tiedonkeruusta ja käytöstä juontuviin syihin. Tulokset on koottu liitetaulukkoon 1 (liite 1). Taulukossa käytettiin pohjana Wangin ja Strongin (1996) mukailtua viitekehystä (taulukko 1). Laatuominaisuuksia lisättiin taulukkoon kirjallisuuden perusteella. Tulokset on esitetty visuaalisessa muodossa kuviossa 3. Kuviossa näkyvät tiedon laadun syy-seurausketjut, joista seuraa tiedon hyödyntämisen ongelmia. Taulukon ellipsin muotoisissa tekstikentissä ovat tiedon laatuominaisuudet ja laatu-ulottuvuudet, joihin ne kuuluvat. Esimerkiksi ihmisiin, teknologiaan ja prosesseihin liittyvät laatuongelmien syyt on sijoitettu suorakaiteen muotoisiin laatikoihin. Nuolet ilmaisevat laatua alentavan vaikutuksen suuntaa.

3.4.1 Tietojen kirjaaminen

Terveydenhuollossa tiedon tallentaminen potilas- ja muihin tietojärjestelmiin tapahtuu suurelta osin käsin. Terveydenhuollon big data onkin alttiimpaa inhimillisille virheille kuin big data yleisesti. (Sukumar ym., 2015.) Kirjaajien erilaiset käyttäytymismallit (Laine ym., 2015) ja yksilölliset työtavat voivat aiheuttaa systemaattista virhettä (Sukumar ym., 2015). Tahattomat ja tahalliset tallennusvirheet ovat yleisiä (Hoffman, 2014; Laine ym.,

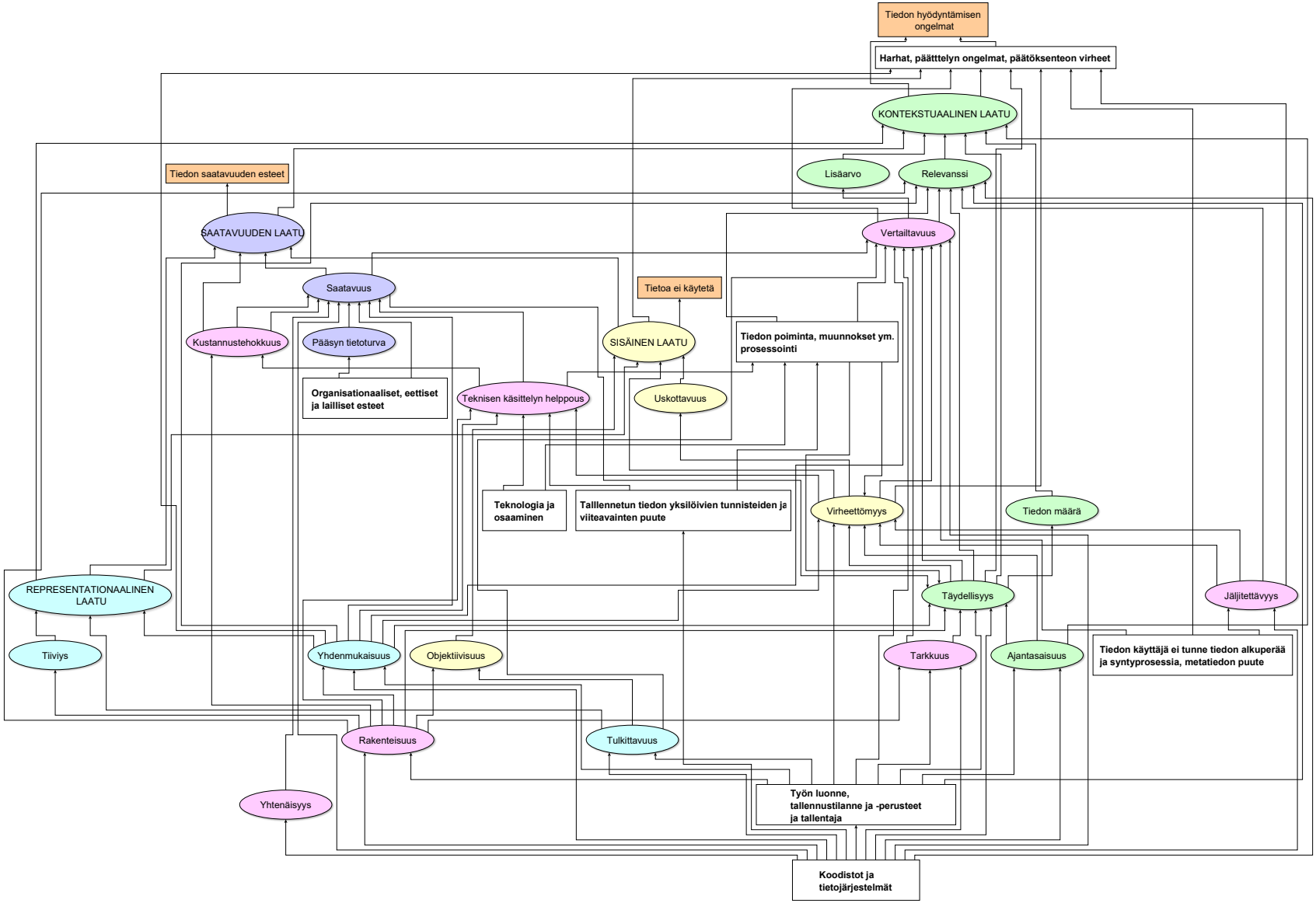
2015; Sukumar ym., 2015; Markus & Topi, 2015). Näitä ovat virheelliset koodit, väärin kirjoitetut sanat ja virheelliset valikkovalinnat (Hoffman, 2014). Joskus tietoja voidaan kirjata väärälle potilaalle, jos useamman potilaan tiedot ovat ruudulla auki (Borycki, 2013).

Myös tiedon puuttuminen on yleistä. Saatavilla olevat tiedot voidaan jättää kirjaamatta jättämällä kenttiä tyhjiksi, kun potilaalla ei ole tiettyjä oireita. Lääkäri voi jättää tietoa myös kokonaan keräämättä, jos potilasta ei kutsuta kontrollikäynnille, ja hoidon teho näin jää toteamatta. (Hoffman, 2014.) Juuri tieto hoidon tuloksista puuttuu erityisen usein (Newgard ym., 2012). Jos lääkäri ei kysy kirjattavista asioista, muista potilaan antamaa tietoa, unohtaa kerrotut asiat tai ei pidä niitä kirjaamisen arvoisina, tärkeääkin tietoa voi jäädä kirjaamatta. (Porter & Mandl, 1999.) Lääkärit myös usein tallentavat tiedon narratiivisessa muodossa, vaikka sille olisi myös rakenteinen tallennusvaihtoehto (Bayley ym., 2013).

Kirjaamisen virheisiin ja puutteisiin on monia syitä, kuten kirjaajien motivaatio, tekninen osaaminen ja kiire. Motiivina esimerkiksi diagnoosin epätarkalle kirjaamiselle voi olla taloudellinen kannustin (Hoffman, 2014; Markus & Topi, 2015). Myös piittaamattomuus ”roskaa sisään, roskaa ulos” -periaatteesta on Molinarin ja Nollon (2020) mukaan yleistä terveydenhuollossa. Kolikon toinen puoli ovat huonot mahdollisuudet keskittyä kirjaamiseen sen vaatimalla tavalla. Terveydenhuollon työympäristö on hyvin kiireinen eivätkä kirjaamisen resurssit ole hyvät (Dungey ym., 2016). Lääkärit usein kirjaavat tietoja vastaanoton aikana ja täydentävät kirjauksia myöhemmin (Bayley ym., 2013). Kirjaaminen vie aikaa ja on ristiriidassa potilaiden hoidon kanssa. Kiireessä tekee helposti näppäilyvirheitä. Tekninen koodaaminen on vaativaa, ja teknisen osaamisen sekä työvoiman puute vaikuttaa osaltaan tiedon laatuun. (Dungey ym., 2016.) Eri syistä johtuvia mittaamisvirheitä myös esiintyy (Hoffman & Podgurski, 2013; Bayley ym., 2013). Muun muassa tekoälysovellusten, kuten koneoppimisen, käyttö voi aiheuttaa systemaattista virhettä (Rajkomar ym., 2018).

Kirjaamisperusteilla on vaikutusta siihen, mikä tietojen laatu on. Usein tietoa kerätään laskutustarkoituksiin (Hersh ym., 2013). Kliininen fokus (Weiskopf & Weng, 2013) ja se, miten lääkäri priorisoi työssään eri kirjaamisperusteita: hoitoa, raportointia ja tutkimusta, vaikuttaa tiedon laatuun suhteessa sen käyttökonteksteihin (Dungey ym., 2016). Paikalliset hallinnolliset ja kliiniset dokumentaatiokäytännöt voivat vaikuttaa paljonkin erityisesti tiedon täydellisyyteen (Nobles ym., 2015). Kun tietoa kerätään tutkimukseen, jossa tietoja aggregoidaan, yhdenmukaisuuden vaatimukset ovat suuremmat, kuin jos tietoa tarkastellaan vastaanotolla potilas kerrallaan. Tutkijan kannalta on myös harmillista, jos lääkäri kirjaa potilaan tietoja vain minimaalisesti, vaikka motiivina olisikin ehkäistä potilaan leimaaminen tai muut kielteiset seuraukset, joita kirjatuista tiedoista voi olla. (Dungey ym., 2016.) Usein syynä tallennetun tiedon huonoon laatuun on se, että kirjaaja ei tiedä tiedon käyttötarkoituksia, eikä osaa ottaa niitä huomioon (Markus & Topi, 2015).

Puuttuva tai virheellinen tieto voi riippua potilaan sairaudesta, lääkärin ammattitaidosta tai siitä, onko henkilö hoidon piirissä. Vaikka kirjauksen perusteena oleva diagnoosi on tärkeää tietoa, sen tekeminen sairauden alkuvaiheessa voi olla vaikeaa (Dungey ym., 2016) tai tauti voi olla vaikeasti diagnosoitavissa. Diagnoosi voi puuttua myös siitä syystä, että henkilö asuu syrjäseudulla, eikä hänellä ole pääsyä erikoislääkärille. (Piri, 2020.) Myös lääkärin ja potilaan huono vuorovaikutus laskee tietojen laatua (Porter & Mandl, 1999).



KUVIO 3 Tiedon laatuongelmat ja niiden syyt terveydenhuollon tiedon toissijaisessa käytössä käyttäen Wangin ja Strongin (1996) viitekehystä.
Ellipsi: laatuominaisuus, jonka laatu alenee, tai jonka heikkous alentaa laatua Vihreä: kontekstuaalinen laatu Keltainen: sisäinen laatu Turkoosi: representationaalinen laatu Violetti: saatavuuden laatu Vaaleanpunainen: Wangin ja Strongin (1996) viitekehykseen kuulumattomat ominaisuudet, Nuoli: tiedon laatua alentavan vaikutuksen suunta

Tietojärjestelmät ja niiden moninaisuus vaikuttaa tiedon laatuun. Erilaisten sairauskertomusjärjestelmien keräämässä tietosisällössä on sekä eroja että yhteneväisyyksiä. Palveluntarjoajat keräävät samoja tietoja korvausvaatimusten ja terveydenhuollon prosessien vuoksi. Näitä tietoja ovat potilaiden demografiset tiedot, terveydenhuollon käynnit, diagnoosit, toimenpiteet, laboratoriotulokset ja vitaaliparametrit. (Richesson ym., 2014.) Kuitenkin sairauden puhkeamisen todellinen ajankohta on vain harvoin nähtävissä potilastietojärjestelmästä (Bayley ym., 2013). Tyypillinen sähköinen potilaskertomus ei myöskään sisällä tietoa taudin vakavuusasteesta ja toimintakyvystä, mikä olisi välttämätöntä tutkittaessa hoidon tuloksellisuutta (Kane, 1997; Richesson ym., 2014 mukaan). Myös lääkärin piirtämät kuvat potilaan sairaudesta ja potilaan itse raportoimat tiedot terveyskäyttäytymisestään ja psykososiaalisista ongelmista puuttuvat sähköisistä potilaskertomusjärjestelmistä (Hoffman, 2014; Estabrooks ym., 2012).

Terveydenhuollon tietojärjestelmät ovat palveluntarjoajakeskeisiä ja vaikka niihin kerätään osin samoja tietoja, tiedot poikkeavat toisistaan tyyppien ja rakenteen osalta (Rayner ym., 2020). Järjestelmissä käytetään erilaisia koodistoja ja tietomuotoja (Jetley & Zhang, 2019), jotka kehittyvät ja muuttuvat myös saman järjestelmän sisällä (Sukumar ym., 2015). Standardoinnin (Dentler ym., 2014) ja semanttisen yhteentoimivuuden puutteesta koituu ongelmia (Liaw ym., 2013; Dungey ym., 2016; Hoffman, 2014; Kruse ym., 2016). Jopa samassa sairaalassa voi olla järjestelmiä, jotka eivät keskustele potilaskertomuksen pääjärjestelmän kanssa (Raghupathi & Raghupathi, 2014b). Näin potilaan kirjaukset sijaitsevat hajallaan (Hoffman, 2014). Tämä hidastaa tiedon kulkua ja työntekoa sekä lisää tarvetta kirjata samoja tietoja moneen paikkaan, mikä altistaa virheille (Hyppönen ym., 2018).

Tietojärjestelmien helppokäyttöisyys, esimerkiksi tyhjien kenttien salliminen, on tiedon laatuvirheiden lähde (Sukumar ym., 2015). Jos tiedot voi tallentaa monella tavalla, lääkäri valitsee usein tavan, joka on helpoin, ja usein tallentaminen narratiivisessa muodossa on helpompaa kuin koodaaminen (Bayley ym., 2013). Myös automaattinen tekstinsyöttö, puheen muuttaminen tekstiksi ja optinen hahmontunnistus voivat aiheuttaa satunnaisia ja systemaattisia virheitä (Sukumar ym., 2015). Ohjelmisto- ja ohjelmointivirheistä voi aiheutua vaikeasti havaittavia ja seurauksiltaan vakavia tiedon laatuongelmia (Hoffman, 2014).

Jo tietojen kirjaamisen tapa, muoto ja järjestelmät sekä kirjaaja ja kirjaamistilanne aiheuttavat siis monia tiedon laatuongelmia. Seuraavaksi siirrymme tarkastelemaan tiedon käyttökontekstin vaikutusta tiedon laatuun.

3.4.2 Toissijainen käyttö

Laadun heikkoudet ovat luonteenomaisia terveydenhuollon tietojärjestelmiin tallennetulle tiedolle. Terveydenhuollon big data koostuu epäyhtenäisistä, epätäydellisistä ja epätarkoista havainnoista, kuten diagnooseista ja hoitotiedoista (Dinov, 2016). Tätä tietoa käytettäessä tiedon todenmukaisuuden puute aiheuttaa suurimmat haasteet (Molinari & Nollo, 2020).

Terveydenhuollon tieto on muita aloja heterogeenisempaa (Jee & Kim, 2013): se on lähtöisin monista lähteistä (Dinov, 2016; Jee & Kim, 2013), rakenteeltaan (Kruse ym., 2016), tietotyypeiltään (Mehta & Pandit, 2018) sekä syntyperältään ja -konteksteiltaan (Weiskopf & Weng, 2013; Laine ym., 2015) ja laadultaan huomattavan vaihtelevaa (Dinov, 2016;

Weiskopf & Weng, 2013; Hersh ym., 2013). Kun erilaisia lähteitä ja tietoja on paljon, laatuongelmat kärjistyvät (Molinari & Nollo, 2020). Riskinä on myös, että aineiston kokonaislaatu laskee huonolaatuisimman osa-aineiston tasolle (Sukumar ym., 2015). Weiskopfin ja Wengin (2013) mukaan erilaiset sähköiset sairauskertomukset poikkeavat mittaamisen, tallentamisen, tietojärjestelmien ja kliinisen fokuksen osalta niin paljon toisistaan, että tiedon laatu on niissä hyvin vaihtelevaa. Siksi tiedon laatua on arvioitava jokaisen aineiston ja käyttötarkoituksen osalta erikseen. (Weiskopf & Weng, 2013).

Tiedon relevanssin puute toissijaisessa käytössä on keskeinen terveydenhuollon tiedon ongelma (Sukumar ym., 2015). Relevanssin puute voi aiheutua siitä, että tieto on kerätty toiseen tarkoitukseen, kuin mihin sitä halutaan käyttää. Taloutta, kuten korvauksia ja laskutusta varten kerättyä tietoa on paljon saatavilla, mutta se ei yleensä täytä kliinisen tai epidemiologisen tutkimuksen vaatimuksia (Sukumar ym., 2015). Tietosisällössä voi olla sellaisia puutteita tai epätarkkuuksia, että tietoa ei voida käyttää. Koodauksen puutteet ovat useiden tutkimusten mukaan erityinen sähköisen potilaskertomuksen käytön este toissijaisessa käytössä (Liaw ym., 2011). Myös koodistoilla on rajoitteita. Esimerkiksi ICD-9 -diagnoosikoodisto ei erottele syöpäpotilaiden alkuperäisiä ja metastaattisia kasvaimia (Botsis ym., 2010).

Tiedon relevanssia tiettyyn käyttötarkoitukseen voi olla vaikea arvioida, sillä on yleistä, että metatieto keruukontekstista puuttuu (Clarke, 2016; Laine ym., 2015). Tästä huolimatta tiedon laatua pidetään usein itsestäänselvyytenä. Esimerkiksi tutkijat, analytiikan toimittajat ja ohjelmistokehittäjät luottavat siihen, että tiedon tarjoava organisaatio on varmistanut tiedon laadun. (Sukumar, 2015.) Heillä ei välttämättä ole käsitystä tiedon alkuperästä ja siitä, millaisen prosessin myötä se on syntynyt. Tämän takia heidän ymmärryksensä tiedon edustavuudesta, laadusta ja harhoista on usein hyvin rajallinen. (Markus & Topi, 2015.)

Terveydenhuollon tiedon standardoinnin puute ja moninaisten tietolähteiden käyttö aiheuttaa terveydenhuollossa väistämättä yhdenmukaisuuden ongelmia. Tietoa on vaikea jakaa organisaatioiden välillä, hankkia, puhdistaa, aggregoida ja analysoida (Kruse ym., 2016). Hoffman (2014) lukeekin yhdeksi puuttuvan tiedon syyksi tietojen harmonisaation puutteen. Täydellinen yhdenmukaisuus on kuitenkin käytännössä mahdottomuus. Dinovin (2016) mukaan terveydenhuollon big data ei voi olla yhdenmukaista ja täydellistä yhtä aikaa. Mikäli halutaan tehdä laajoja, esimerkiksi kohortti- tai väestötason analyyseja kattavammalla aineistolla, tieto on vähemmän yhdenmukaista, kuin jos analysoidaan vain rajattua joukkoa, jossa tiedot voivat vastaavasti olla yhdenmukaisempia (Dinov, 2016). Sekä tiedon yhdenmukaistaminen että yhdenmukaisuuden puute voivat siis johtaa paitsi aineiston käsittelyn työläyteen myös aineiston epätäydellisyyteen.

Terveydenhuollon tiedon laatua toissijaisessa käytössä vähentää puuttuva tieto (Weiskopf & Weng, 2013). Sairauskertomusaineistot sisältävät keruu- ja varastointimenetelmiensä ja rakenteensa vuoksi laajaa systemaattista harhaa, eikä niiden täydellisyys ole pelkästään puuttuvan tiedon satunnaisuuden asteen perusteella määriteltävissä. Puuttuva tieto voi olla sekä aineiston sisäistä, käyttökontekstista riippumatonta, että ulkoista, tiettyyn käyttökontekstiin liittyvää (Weiskopf ym., 2013b, taulukko 4.) Hoffman (2014) puhuu puuttuvasta tiedosta tietotyhjiöinä. Tällaisia tyhjiöitä voi olla havaintoyksiköiden tai muuttujien tasolla. Potilaiden historian epätäydellisyys ja tietojen yksityiskohtaisuuden puute (Dentler ym., 2014), toisin sanoen puuttuvat attribuutit, puuttuva tieto niiden sisällä ja havaintoaineiston rajoittaminen (Jetley & Zhang, 2019), vähentävät tiedon käyttökelpoi-

suutta. Toissijaisessa käytössä keskeinen tiedon ulkoinen tyhjiö koskee terveitä ihmisiä, joiden terveystiedot ovat sairaita ihmisiä puutteellisemmat, miksi potilaskertomustiedot eivät ole väestötasolla edustavia (Weiskopf ym., 2013; Hoffman, 2014). Suuri osa tietotyhjiöistä tulee toissijaisen käytön näkökulmasta annettuna, eikä niihin voi enää vaikuttaa, jos tietoa ei ole syystä tai toisesta tietojärjestelmiin kerätty (esim. Estabrooks ym., 2012; Laine ym., 2015; Weiskopf ym., 2013b).

Myös tietojen virheellisyys on keskeinen potilastiedon laatuongelma (esim. Dentler ym., 2014). Tiedon paikkansapitämättömyys voi johtua sekä tiedon sisäisistä systemaattisista ja satunnaisista virheistä että tiedon eri formaateista ja merkityksistä eri tallennus- ja käyttökonteksteissa (taulukko 5). Laine ym. (2015) tutkivat aikaleimojen tallennusta sairaalan tietojärjestelmiin. Tiedon laatuongelmia aiheutui epäselvistä tai epäyhtenäisistä määritelmistä, jolloin sama aika saattoi olla tallennettu hallinnollisten ohjeiden mukaan oikein, mutta se ei vastannut todellisuutta sairaalan prosesseista. (Laine ym., 2015.) Myös Goldberg, Niemierko, & Turchin (2008) havaitsivat tutkimuksessaan, että sisäisestä epäyhdenmukaisuudesta johtuvat virheet terveydenhuollon tietokannoissa eivät usein ole satunnaisia, eikä niistä löydetä kuin murto-osa käyttämällä sääntöpohjaisia rajoituksia. Sisäisestä epäyhdenmukaisuudesta aiheutuu ongelmia paitsi prosessien louhinnassa (Laine ym., 2015), myös tutkimuksessa. Jos tietojen tallennustavoista eri aikoina eri paikoissa ei ole tietoa tai tietoja ei ole tallennettu yhdenmukaisella tavalla, tietojen vertailu tai erojen tulkitseminen muutokseksi on mahdotonta (Bayley ym., 2013).

TAULUKKO 5 Tiedon paikkansapitämättömyyden luokittelu (Laine ym., 2015)

Varsinaiset virheet		Semanttinen heterogeenisuus	
Satunnaiset virheet	Systemaattiset virheet	Representationaalinen heterogeenisuus	Ontologinen heterogeenisuus
”Virheellinen tieto”		”Kontekstien yhteensopimattomuus”	

Yhdenmukaisuuden puute on siis keskeinen terveydenhuollon tiedon laatuongelma, sillä se vaikuttaa alentavasti sekä tiedon virheettömyyteen (Laine ym., 2015; Hoffman & Podgurski, 2013) että täydellisyyteen (Hoffman, 2014) tiedon toissijaisessa käytössä. Tallentamistapojen erilaisuudesta aiheutuu myös vertailtavuuden ongelmia. Chanin, Fowles'in ja Weinerin (2010) määritelmän mukaan tiedon vertailtavuus on vertailussa käytettävien dataelementtien saatavuuden ja laadun samanlaisuutta. Yhdenmukaisuuden vaatimus hoidon laatua vertailevassa tai hoidon tuloksellisuutta arvioivassa tutkimuksessa koskee siis paitsi tiedon sisällön, formaatin ja rakenteen yhdenmukaisuutta, myös sitä, että vertailtavan tiedon tulisi olla yhtä hyvin saatavilla ja täydellisyydeltään ja virheettömyydeltään samanlaista.

Tiedon rakenne on Krusen (2016) kirjallisuuskatsauksen mukaan suurin haaste terveydenhuollon big datan hyödyntämisessä. Rakenteeton tieto on erityisen epäyhdenmukaista: se on hyvin vaihtelevaa ja epätarkkaa (Raghupathi & Raghupathi, 2014b). Rakenteettomalla tiedolla useimmiten viitataan narratiiviseen, vapaaseen potilaskertomustekstiin, joka sisältää samoja asioita monin eri tavoin ilmaistuna. Eri lääkärit voivat esimerkiksi käyttää samoja lyhenteitä eri asioita tarkoittaessaan. (Hoffman, 2014.) Narratiivista tekstiä käytetään, koska usein muutakaan tiedon lähdettä ei ole tarjolla (Bayley ym., 2013). Sitä voidaan myös pitää täysin käyttökelvottomana (Dentler ym., 2014). Rakenteisen tiedon puute on sitä hankalampi ongelma, mitä keskeisemmästä tiedosta on kyse (Bayley ym.,

2013). Rakenteettomat aineistot ovat erityinen ongelma myös organisaatioiden sisäisesti, koska niitä on vaikeaa aggregoida ja analysoida (Kruse ym., 2016).

Sen lisäksi, että jo alkuperäisessä tiedossa on paljon laadun heikkouksia käyttökontekstin näkökulmasta, sen prosessoiminen aiheuttaa niitä. Tiedon kontekstuaalinen heterogeenisuus ja laatuongelmat vaativat paljon manuaalista prosessointia (Laine ym., 2015). Inhimillisistä virheistä ja yksilöllisistä työtavoista johtuvat tiedon laatuongelmat ovat yleisiä tiedon käytössä. Tiedon yhdenmukaistaminen, erityisesti useat ETL-prosessit lisäävät laatuvirheiden mahdollisuutta. (Sukumar ym., 2015.) Tietoihin tehdyt muunnokset voivat heikentää laatua tietyn käyttökontekstin näkökulmasta (Hersh ym., 2013). Ad hoc -tyyppinen puuttuvan tiedon käsittely, esimerkiksi puuttuvien arvojen interpolointi (Clarke, 2016), on riskialtista.

Monista tietolähteistä seuraa vaikeita laatuongelmia erityisesti tietorakenteiden ja -tyyppien yhdistämisessä (entity matching) (Molinari & Nollo, 2020). Terveydenhuollon tiedossa on paljon erilaisia entiteettejä, kuten palveluntarjoajia, potilaita, maksajia ja sääntelyviranomaisia (Sukumar ym., 2015), mutta usein yksilöivät tunnisteet (Molinari & Nollo, 2020) ja tietoalkioiden väliset yhteydet (Dentler ym., 2014) puuttuvat. Tiedon yhdenmukaistamisvaiheessa riskinä on entiteettien hajoaminen (Sukumar ym., 2015). Koska potilaiden kirjaukset ovat hajallaan eri lähteissä, on mahdollista poimia sama potilas aineistoon monta kertaa, erityisesti, jos aineistosta on poistettu yksilöivät tunnisteet (Hoffman, 2014). Riskinä on myös yhdistää kaksi lähes identtistä tapahtumaa (Clarke, 2016) ja se, että tieto lähteestä hämärtyy. Lähdetiedon epäselvyys on riski erityisesti lääkitystiedon osalta, koska sillä on yleensä useita lähteitä samankin organisaation sisällä (Richesson ym., 2014).

Kliinisen hoidon prosessi tuottaa sairauskertomusaineistoihin laajaa systemaattista harhaa, jollaista ei ole perinteisissä tutkimusaineistoissa (Weiskopf ym., 2013b). Epäkokeellisessa potilaskertomustietoa hyödyntävässä tutkimuksessa vakavia laatuongelmia aiheutuukin aineistojen systemaattisesta harhasta ja havaitsemattomista sekoittavista tekijöistä (Hoffman & Podgurski, 2013). Hoffman ja Podgurski (2013) mainitsevat erityisesti valikoitumisharhan, sekoittumisharhan ja mittaamisharhan, joiden huomiotta jättäminen tai väärä tulkinta johtaa virheellisiin tuloksiin (Hoffman & Podgurski, 2013). Hyvinkin pienellä virheellisten tietojen osuudella voi olla suuri vaikutus tutkimuksen tuloksiin (Hripcsak ym., 2011). Erityisen ongelmallista on, että näennäistieteellisiä ja harhaanjohtavia tuloksia voidaan käyttää poliittisten, sosiaalisten ja taloudellisten tarkoitusperien ajamiseen (Hoffman & Podgurski, 2013). Ne voivat myös johtaa vakaviin seurauksiin yksilöille ja ihmisryhmille, kun niitä käytetään päätöksenteon perusteena (Clarke, 2016).

Sähköisten sairauskertomusjärjestelmien rakenne ja ominaisuudet ovat kriittinen tekijä ja usein myös tuntematon sekoittava tekijä terveydenhuollon järjestelmien ja kliinisen tiedon tutkimuskäytössä (Holve ym., 2013). Tiedon prosessoinnin teknologioista ja menetelmistä aiheutuu laatuongelmia, joita voi olla hankala kontrolloida. Ne voivat liittyä esimerkiksi laitteistoalustojen välisten standardien noudattamatta jättämiseen, vaihtelevaan tiedon saatavuuteen, tiedon päivittämisen vaihtelevaan nopeuteen tai järjestelmien häiriöihin. Tiedon käsittelyn, varastoinnin ja prosessoinnin automatisoiminen voi paitsi helpottaa työtä myös aiheuttaa laatuongelmia. Erityisesti terveydenhuollossa tiedon laatuongelmat voivat olla niin spesifejä, että on rakennettava räätälöityjä sovelluksia ja omia välineitä tiedon laadun sääntöpohjaiseen varmistamiseen. (Sukumar ym., 2015.)

Monet analyysimenetelmät vaativat tiedolta hyvin korkeaa laatua tuottaakseen laadukkaita tuloksia. Siksi koneoppimisen metodien ja analyyttisten algoritmien käyttö ai-

heuttaa erityisiä haasteita terveydenhuollossa (Sarafidis ym., 2020; Sukumar ym., 2015). Big data-analytiikan käyttö terveydenhuollon päätöksenteossa sisältää suuria riskejä (Clarke, 2016). Menetelmät ja välineet kuitenkin kehittyvät jatkuvasti. Luonnollisen kielen prosessoinnin tekniikoiden käyttö vapaan tekstin automaattisessa luokittelussa on vaatinut aiemmin paljon manuaalista työtä, mutta uudemmat, enemmän laskentatehoa vaativat syväoppimistekniikat ovat helppokäyttöisempiä (Chen ym., 2018).

Terveydenhuollossa tiedon käytön haasteita aiheutuu myös taloudellisten resurssien, sopivien teknologioiden, välineiden, osaamisen ja tieteellisen näytön puutteesta. Terveydenhuollon organisaatioiden on vaikea löytää niiden omaan käyttöön sopivia big data -teknologioita avoimen lähdekoodin (esim. Hadoop) ja kaupallisten vaihtoehtojen (esim. Cassandra) joukosta (Jee & Kim, 2013). Tietoturvan kysymykset ovat terveydenhuollossa erityisen kriittisiä, mutta etenkään avoimen lähdekoodin big data -teknologiat eivät tarjoa riittäviä välineitä tietoturvaan (Jee & Kim, 2013; Raghupathi & Raghupathi, 2016). Lisäksi osaavien analyytikoiden tai datatieteilijöiden rekrytointi on hankalaa ja kallista (Mehta & Pandit, 2018; Jee & Kim, 2013). Tietovarastoinnin (Kruse ym., 2016) ja analytiikkatyökalujen (Mehta & Pandit, 2018) kustannukset ovat usein esteenä big datan käytölle. Big data -analytiikkaan ei olla valmiita panostamaan ennen kuin sen hyödyistä saadaan riittävästi tieteellistä näyttöä (Mehta & Pandit, 2018).

Organisaationaaliset, eettiset ja lailliset esteet saattavat olla teknologisiakin esteitä suuremmat, esimerkiksi kun halutaan mitata lääketieteellisten hoitojen vaikuttavuutta laajassa mittakaavassa (Molinari & Nollo, 2020). Terveydenhuollon tiedot ovat siiloutuneina, erillään ja tietosuojasäädösten vahvasti suojaamina palveluntarjoajien ja viranomaisten rekistereissä (Jee & Kim, 2013). Tiedon omistajuus on erittäin sirpaloitunut. Omistajia ovat esimerkiksi terveyspalvelujen tarjoajat, vakuutusyhtiöt, lääkeyritykset, valtio, kolmannen osapuolen tiedonvälittäjät, potilasasiakirjajärjestelmien (EMR) tuottajat ja potilaat itse. (Szlezák ym., 2014.) Potilastiedot ovat arkaluontoista tietoa, jonka jakaminen ja käyttö edellyttää vahvoja tietoturva- ja kompliansissisäädöksiä (Jee & Kim, 2013). Kansalliset yksityisyyssäännökset ja se, että asiakkaat voivat määrätä omien tietojensa käytöstä ja säilyttämisestä, vaikeuttaa tiedon käyttöä (Buhl ym., 2013).

Ilman ylhäältä alas -lähestymistapaa tietosiilojen purkaminen ja big datan tehokas hallinta ja integroiminen on mahdotonta (Jee & Kim, 2013). Suomessa tiedon käytön esteitä onkin purettu yhtenäistämällä hajanaisia järjestelmiä kansallisin sosiaali- ja terveysalan digitalisaatiohankkein ja lainsäädäntöä uudistamalla (Hyppönen & Ilmarinen, 2016). Keskeinen uudistus on vuonna 2019 voimaan tullut Laki sosiaali- ja terveystietojen toissijaisesta käytöstä, niin sanottu toisiolaki. Se helpottaa tietojen käyttöä muun muassa tutkimukseen, kehittämiseen ja tietojohtamiseen (tiedolla johtamiseen). (STM, 2019; THL 2020.) Suomessa on myös luotu kansallisia terveyden ja hyvinvoinnin tietovarantoja, kuten Kansaneläkelaitoksen Kanta-palvelujen ylläpitämä Potilastiedon arkisto (Lääkäriliitto, 2020). Lisäksi on pyritty resursoimaan toimijoita tiedon hyödyntämisen vaatimissa uudistuksissa, kuten tietoaltaiden perustamisessa (Sitra, 2018). Kehittämistä on vielä muun muassa tiedon löydettävyydessä, tietoon pääsyssä ja siinä, että analysoinnin kustannukset eivät nouse liian korkeiksi yksittäisille tutkimusryhmille (Darst, Hakala & Kaski, 2018).

Terveydenhuollon tiedon toissijaisen käytön esteet ovat siis hyvin moninaisia ja laaja-alaisia. Sen toissijaisessa käytössä on tärkeää huomioida paitsi tiedon tekniset laatuominaisuudet myös tiedon keräämisen ja käytön kontekstit sekä koko se sosiotekninen prosessi, jossa tieto on syntynyt ja jossa sitä on jalostettu. Jotta terveydenhuollon big dataa

voidaan hyödyntää, se edellyttää tiedon käyttäjille parempaa pääsyä siiloutuneisiin tiedon lähteisiin, jotka ovat tietosuojasäännösten vahvasti suojaamia. Terveydenhuollon organisaatiot tarvitsevat myös parempia taloudellisia, teknologisia ja inhimillisiä resursseja sekä tietoa big data -analytiikan käytännön hyödyistä voidakseen jalostaa big datasta arvoa. Kattavat tiedon laatuun liittyvät vastuut, prosessit ja menettelytavat sekä tiedon jäljitettävyys metadatan avulla varmistavat tiedon riittävän laadun ja ovat sen edellytys terveydenhuollon tiedon toissijaisessa käytössä (Clarke, 2016). Tiedon laatuun vaikuttavat myös laajemmat, alan sisäiset, yhteiskunnalliset ja lainsäädännölliset tekijät, kuten tutkimusrahoitus, tietosuojasäädökset, lääketieteelliset koodistot ja laitteiden standardit.

3.5 Yhteenveto

Big data määritellään usein määrältään runsaaksi, nopeasti kertyväksi ja moninaiseksi tiedoksi, jonka todenmukaisuus (laatu) on alhainen ja josta on mahdollista jalostaa runsaasti arvoa (esim. Baesens ym., 2016). Big dataa kertyy esimerkiksi organisaatioiden suurista tietojärjestelmistä, esineiden internetistä (esim. terveyssensorit), internetin klikkausvirrasta ja sosiaalisesta mediasta (Baesens ym., 2016). Terveydenhuollolle ominaisia big datan lähteitä ovat muun muassa sairauskertomukset, laboratoriotulokset ja geneettinen informaatio (Sukumar ym., 2015). Big datan laatu ja inhimilliset tekijät sen hyödyntämisessä ovat jääneet tutkimuksessa vielä paljolti huomiotta (esim. Baesens, 2016; Mikalef ym., 2018).

Tiedon laatukysymykset ovat tärkeitä, sillä ilman hyvälaatuista tietoa, koituu säästöjen ja hyötyjen sijaan kustannuksia, ja esimerkiksi terveyspalveluntuottajan kilpailukyky alenee. Ennen kaikkea huono tiedon laatu terveydenhuollon päätöksenteon pohjana on riski potilaiden terveydelle. Big datan yhteydessä tällaiset riskit kasvavat perinteiseen dataan verrattuna (Clarke, 2016).

Yleinen hyvälaatuisen tiedon määritelmä on, että tieto sopii käyttötarkoitukseensa (fitness for use) (Strong ym., 1997). Strongin ym. (1997) mukaan tällainen tieto on teknisesti laadukasta, uskottavaa ja maineeltaan hyvää sekä esittämistavaltaan, saavutettavuudeltaan laadukasta ja sopii käyttökontekstiinsa. Tiedon laatuominaisuuksia ovat kirjallisuudessa esimerkiksi täydellisyys, virheettömyys ja ajantasaisuus, mutta erilaiset laatuluokitukset poikkeavat toisistaan paljonkin (Laranjeiro ym., 2015). Tiedon laatuongelma on mikä tahansa, missä tahansa laatu-ulottuvuudessa kohdattu ongelma, joka tekee tiedosta kokonaan tai suurelta osin käyttöön sopimattoman. (Strong ym., 1997.) Tiedon huono laatu on monimutkaisten organisationaalisten prosessien tulosta (Strong ym., 1997), ja laatuongelmat voivat liittyä mihin tahansa tiedon elinkaaren vaiheeseen (Sukumar ym., 2015).

Big datan hyödyntämisen kannalta on olennaista, miten laadukasta organisaatioiden omien suurten tietojärjestelmien sisältämä tieto on. Terveydenhuollossa sähköiset sairauskertomukset ovat keskeinen tiedon lähde, johon halutaan liittää muita aineistoja ja käyttää näin syntyneitä yhdistelmäaineistoja esimerkiksi hallinnolliseen päätöksentekoon ja tutkimuksellisiin tarkoituksiin (Weiskopf & Weng, 2013). Terveydenhuollon tiedossa on erityisiä laatuhaasteita, jotka liittyvät muun muassa tiedon relevanssiin sen käyttökonteksteissa, manuaaliseen tiedonsyöttöön, moninaisiin tiedon lähteisiin ja muuttuviin standardeihin (Sukumar ym., 2015). Myös tietoturvaan ja yksityisyyteen liittyvät kysymykset vaikuttavat terveystiedon laatuun (Buhl ym., 2013).

4 TAPAUSTUTKIMUKSEN KONTEKSTI JA TOTEUTUS

Tässä luvussa kuvataan tutkimuksen tapaus ja konteksti, empiirisen aineiston keruu- ja analyysimenetelmät sekä -prosessi. Lisäksi käsitellään tutkimuksen luotettavuutta. Tutkijan suunnitteluvaiheessa hankkimaa esitietoa kontekstista on kuvattu tämän luvun alkuosassa. Luvun loppuosassa kuvaillaan vielä urologian klinikan tietoallashanketta haastattelujen ja dokumentaation perusteella.

4.1 Tutkimuksen tapaus ja konteksti

Tutkija kartoitti keväällä 2019 tutkimukseen sopivaa tapausta keskusteluissa eri sairaanhoitopiirien ja muiden tahojen asiantuntijoiden kanssa. Tarpeeksi pitkälle edenneitä terveydenhuollon big dataa hyödyntäviä hankkeita oli vaikea löytää. VSSHP:n Auria Tietopalvelun tietopalvelujohtaja Arho Virkki ehdotti tutkimukseen vuonna 2016 alkanutta ”Urologian hoitopolku” -hanketta, josta oli jo kertynyt kokemuksia noin kolmen vuoden ajalta. Kun tutkija lisäksi sai tutkimukselle tukea VSSHP:n Auria Tietopalvelusta, tutkimuksen toteuttaminen tästä hankkeesta mahdollistui. Haastatteluvaiheessa tapaus laajeni koskemaan kokemuksia VSSHP:n tietoaltaan tiedon käytöstä yleisemmin tutkimukseen ja tiedolla johtamiseen (ks. luku 4.3.4). Konkreettisten tiedon laatuongelmien ratkaisujen osalta urologian hanke toimi ”upotettuna” tapauksena (ks. Yin, 2018).

Tutkimuksen kontekstina oli VSSHP:n tietoallastiedon jalostamiseen rakennettu ja muodostunut sosiotekninen kokonaisuus, joka sisältää ihmiset ja organisaation, tiedonkäsittelyprosessit, vuorovaikutuksen, tiedon, teknologiainfrastruktuurin ja tietojärjestelmät. Tiedon laatuongelmia ja niiden syntyprosesseja kuvattiin tässä ympäristössä. Seuraavassa kuvataan tutkijan alustavaa esitietoa VSSHP:n tietoaltaasta, Auria Tietopalvelusta, tietoallastiedon prosessoinnista sekä urologian klinikalle kehitetystä aikajananäkymästä. Esitietoa hankittiin pääasiassa julkisesti saatavilla olevista dokumenteista ja hankkeen avainhenkilöiden kanssa käydyissä keskusteluissa (ks. luku 4.3.1).

4.1.1 Varsinais-Suomen sairaanhoitopiirin tietoallas

VSSHP:n tietoallas rakennettiin Isaacus-esituotantohankkeena vuosina 2016–2017(Sitra, 2018). Isaacus-esituotantohankkeet valmistelivat kansallisen tietolupaviranomaisen, vuoden 2020 alussa aloittaneen Findatan, toimintaa (STM, 2019; THL 2020), ja niiden tuloksena syntyivät VSSHP:n, Helsingin ja Uudenmaan sekä Pohjois-Savon sairaanhoitopiirin tietoaltaat ja niiden ympärille tietotaitoa tietoallasteknologioista, tiedon laadusta ja hallinnasta. Lisäksi esimerkiksi VSSHP:n kliinisen tietopalvelun kyvykkyyksiä kehitettiin visuaalisessa analytiikassa ja data-analyysissä. (Sitra, 2018.)

VSSHP:n tietoallas kattaa TYKS:in erityisvastuualueen (Varsinais-Suomen, Satakunnan ja Vaasan sairaanhoitopiiri) ja sen tuottaa 2M-IT (entinen Medbit Oy). Se on toteutettu Hadoopin hajautetulla levyjärjestelmällä. Ratkaisuna on Cloudera ja se sijaitsee 2M-IT:n konesalissa (Darst, Hakala & Kaski, 2018). Tiedosta voidaan tutkia annettuja hoitoja, hoidon laatua ja vaikuttavuutta, kustannusrakennetta ja tieteellisiä kysymyksiä (Virkki, 2017). Tietoaltaan keskeisimmät lähdejärjestelmät on lueteltu taulukossa 6 (Virkki (2017), tuottajat-sarakkeen tiedot: 2M-IT:n palvelupäällikkö Ari Wahlstedt).

Saatavilla oleva tieto vaihtelee lähdejärjestelmän mukaan (ks. Auria Tietopalvelu, 2020a). Kattavaa tietoa on noin vuosikymmenen ajalta. Keskeisimpien potilasrekisteritietokantojen lisäksi altaassa on tietoa pienemmistä potilastietojärjestelmistä, BCB Medicalin laaturekisteritietoa ja esimerkiksi potilaskyselyjen aineistoa. Tällä hetkellä altaassa on pääasiassa vain erityissairaanhoidon tietoja. (Virkki, 2017.) Perusterveydenhuollon tietoja on Salosta ja Paimio-Sauvosta (tarkennus haastattelun perusteella L1, ks. taulukko 9).

TAULUKKO 6 VSSHP:n tietoaltaan keskeisimmät tiedon lähteet

Tietojärjestelmä	**Tietoja**	**Tuottaja**
Uranus	Potilastietojärjestelmä	CGI
Oberon (Uranuksen osajärjestelmä)	Syntymä- ja kuolinajat, osastohoidot, avohoidot, lähetteet, toimenpiteet, diagnoosit	CGI
Qpati	Esitiedot, patologin lausunnot ja patologian diagnoosit	Tieto
Opera	Leikkaukset, toimenpiteet ja diagnoosit	GE
Miranda/Desktop (Uranuksen osajärjestelmä)	sairauskertomusteksti, hoitotaulukot, lääkeosio, sähköiset reseptit	CGI
Radu	kuvantamistutkimukset, toimenpidekoodit ja päivämäärät	L-Force
TYKSLAB	laboratoriotutkimukset, tulokset ja viitearvot	My+ (entinen MyLab)
Kemokur (Uranuksen osajärjestelmä)	Sytostaattihoitojärjestelmä	CGI
Aria	Sädehoidon annossuunnittelujärjestelmä (ei sis. brakihoitoja)	Varian
WebMarela	Apteekin toimittamat iv-sytostaatit	AffectoGenimap

4.1.2 Sairaanhoitopiirin tietopalvelu

Auria Tietopalvelu (viralliselta nimeltään Varsinais-Suomen sairaanhoitopiirin tietopalvelu) on vuonna 2014 perustettu VSSHP:n yksikkö, joka jalostaa tietoaltaan tietoa toisiokäyttöön organisaation sisäisten ja ulkoisten käyttäjien tarpeisiin. Tietopalvelun tehtävissä työskentelee 13 eri koulutus- ja kokemustaustaista henkilöä mm. tietoarkkitehtina, it-suunnittelijoina, tilastotieteilijöinä ja tietokantakehittäjinä. Se toimii läheisessä yhteistyössä sairaalan kliinikoiden ja VSSHP:n Auria Biopankin kanssa. Tietopalvelun henkilöstö vastaa muun muassa tietojen mallintamisesta, dokumentaatiosta, puhdistamisesta, poiminnasta, harmonisoinnista (yhteismitallistus), käyttöoikeuksien myöntämisestä ja aineistojen luovuttamisesta. (Auria Tietopalvelu, 2020b; I3, ks. taulukko 9.)

4.1.3 Tietoallastiedon laadun varmistaminen ja yhteismitallistaminen

Keskeisiä vaiheita tietoaltaassa olevan tiedon laadun varmistamisessa ovat tietoallastiedon lataus ja tietojen yhteismitallistaminen eli harmonisointi. Lähdejärjestelmien tietojen ETL-prosessi sisältää kuusi vaihetta, jotka ovat tiedon hakeminen lähdejärjestelmästä, tiedon lataaminen, formaattien konversio, tyyppien konversio, tietojen yhdistäminen ja semanttinen yhtenäistäminen. (Hammais, Varjonen & Virkki, 2018.) Laadunvarmistusta tehdään tietolähdekohtaisesti seuraamalla rajapintojen kautta tapahtuvia päivityksiä, vertaamalla tietoa siitä annettuun kuvaukseen ja tarkistamalla, onko tiedossa virheitä. Tiedon oikeellisuudesta voidaan varmistua yhdistämällä tietoa monesta lähteestä. Esimerkiksi syöpä kliinisenä diagnoosikirjauksena varmentuu patologian tiedoista, toimenpiteistä, kuvantamisesta ja laboratoriotiedoista. Rakenteetonta potilaskertomustekstiä käytetään varmentamaan tietoa, kun rakenteinen tieto on puutteellista. (Virkki, 2017.)

Tietojen yhteismitallistaminen tapahtuu pyynnöstä ja alkaa tiedon profiloinnilla. Siinä selvitetään tiedon omistaja, aineistossa käytetyt koodausjärjestelmät ja tietotyypit ja se, mitä puuttuvia arvoja aineistossa on. Kaikki edellä mainituista voivat vaihdella tiedonkeruuajankohdan mukaan. Seuraavassa vaiheessa yhtenäistetään aineistoissa käytetyt koodistot esimerkiksi muuntamalla paikallisesti käytetyt koodit kansallisen koodiston mukaisiksi. Koodien yhteismitallistamisen jälkeen tiedolle on luotava ymmärrys ja keskusteltava tiedon kirjaajien kanssa, jotta voidaan erottaa erilaiset kirjaamisen tavat ja löytää kaikista relevanteimmat tiedot. Seuraavaksi aineistosta etsitään analyysissä käytettävät muuttujat, niiden taso ja mahdollinen järjestys. Vapaamuotoisesta tekstistä poimitaan tarvittaessa muuttujia. Tämän jälkeen tieto mallinnetaan ja yhteismitallistetaan, ja sitä voi alkaa aggregoida ja käyttää. (Hammais, Varjonen & Virkki, 2018.)

4.1.4 Urologian aikajananäkymä

Tutkimuksen tarkoituksena oli tutkia VSSHP:n tietoaltaan hyödyntämiseen liittyviä tiedon laatuongelmia ”Urologian hoitopolku” -hankkeessa. Hanke aloitettiin osana Sitran osittain rahoittamaa VSSHP:n hanketta ”360 asteen näkymä jalostettuun potilastietoon – sairauskertomustiedon hyödyntäminen johtamisessa ja potilaan hoidossa”. Hanke oli yksi Sitran Isaacus-projektin esituotantohankkeista. (ks. Sitra, 2018.) Se sisältää vuonna 15.6.2016 al-

kaneen pilotin, jossa on kehitetty lääkärille potilaan aikajananäkymän prototyyppi urologian klinikan potilaiden hoidon laadun seurantaan ja kehittämiseen. Tieto kerätään ja yhdistetään yli kymmenestä erilaisesta sähköisen potilastiedon rekisteristä (ks. taulukko 6) ja sisältää kuvantamis-, kemoterapia-, toimenpide-, diagnoosi- ja hoitotietoja, jotka on viety sairaanhoitopiirin tietoaltaaseen. Hankkeessa räätälöidään analyysi- ja visualisointityökaluja tiedon hyödyntämiseen. (Virkki, 2017.) Projekti oli tämän tutkimuksen alkaessa ollut meneillään noin kolme vuotta ja tässä ajassa oli kertynyt kokemuksia tietoallastiedon käytöstä.

4.2 Tiedonkeruumenetelmät ja niiden valinta

Tutkimuksen tutkimusote oli laadullinen. Tutkimusstrategiaksi valittiin tapaustutkimus, jossa tietoa kerättiin puolistrukturoiduin haastatteluin ja perehtymällä julkisiin dokumentteihin. Valinnoissa huomioitiin tutkimusalue, tutkimuskysymykset ja tutkimuksen teon käytäntö sekä tutkijan ja organisaation resurssit.

Big datan laadun sosioteknisten aspektien tutkimus on uusi tutkimusalue, jota koskeva kirjallisuus on niukkaa. Tutkimuskysymyksiin liittyvät informaatioprosessit ja organisaatioilmiöt ovat monimutkaisia, ja niistä saatava tieto riippuu tietoa tuottavien, sitä käsittelevien sekä sitä käyttävien henkilöiden (asiantuntijoiden) subjektiivista kokemuksista. Laadulliset tutkimusmenetelmät sopivat tämänkaltaiseen tutkimukseen: ilmiö on vähän tunnettu, tutkimuskysymykset kompleksisia ja kontekstisidonnaisia (Patton, 2002), ja tutkimus pyrkii syvälliseen tietoon ja ymmärrykseen tutkimuskohteesta sen omassa organisationaalisessa ympäristössä (Carter & Little, 2007; Hoepfl, 1997; Patton, 2002).

Kun tutkitaan tietojärjestelmää tai nykyilmiötä eli ”tapausta” syvällisesti sen tosielämän kontekstissa, tapaustutkimus on hedelmällinen tutkimusote ja -menetelmä (Darke, Shanks, & Broadbent, 1998; Yin, 2018), erityisesti, kun ilmiö ei ole erotettavissa kontekstistaan. Tapaustutkimuksessa käytetään erilaisia tiedon lähteitä, jotka valottavat samoja tutkimuskysymyksiä eri näkökulmista (triangulaatio). Tavoitteena on erityisesti vastata kysymyksiin ”miksi?” ja ”miten?”(Yin, 2018.) Yinin (2018) mukaan tapaustutkimukset ovat yleistettävissä teoreettisiin propositioihin, eivät populaatioihin. Edellä mainitut lähtökohdat toteutuvat tässä tutkimuksessa, jossa yritettiin löytää tietyssä terveydenhuollon tiedon toissijaisen käytön kontekstissa säännönmukaisesti ilmeneviä tiedon laatuongelmia ja niitä sosioteknisiä prosesseja, joissa ne syntyvät. Analyyttisenä tavoitteena oli selittää tiedon toissijaisessa käytössä ilmeneviä tiedon laatuongelmia terveydenhuollon big data -ympäristössä. Tapaustutkimus on ihanteellinen menetelmä tällaiseen tutkimukseen.

Tapaustutkimus yleisesti on vaativa käytännössä (Darke ym., 1998; Yin, 2018). Tapauksen valinta oli haastavaa, koska mahdollisten, teoreettiseen otantaan sopivien tapausten määrä oli hyvin pieni. Vain muutamassa sairaanhoitopiirissä oli käytössä tietoallas ja meneillään sitä hyödyntäviä hankkeita. Monitapaustutkimuksella olisi ollut yksittäistä tapausta paremmat edellytykset tarjota tieteellisesti vahvoja vastauksia (Yin, 2018). Tarjolla olevien tapausten vähyys ja tutkijaresurssit mahdollistivat tässä kuitenkin vain yksittäisen tapauksen (ks. Yin, 2018) käytön. Sekä tutkijan että haastateltavien rajallisten resurssien vuoksi aineistonkeruu oli suunniteltu mahdollisimman tehokkaaksi.

Tapaustutkimuksessa on tärkeää hankkia riittävä pääsy ”tapaukseen” (Yin, 2018). Terveydenhuollon ympäristö on erityisen haastava, sillä suuri osa toiminnasta ja tiedosta on potilaiden yksityisyyteen liittyvää ja arkaluontoista. Tutkijan ei ollut mahdollista päästä havainnoimaan päivittäistä toimintaa tai käsiksi potilastietojärjestelmien tietoon. Tiedonkeruu keskittyi sen sijaan asiantuntijoiden ymmärrykseen tiedosta ja tietoon liittyvistä prosesseista, josta saatiin tutkimusmateriaalia puolistrukturoiduin haastatteluin. Niistä syntyvää kuvaa pohjustettiin, täydennettiin ja varmennettiin julkisten dokumenttien avulla.

Yin (2018) pitää tapaustutkimukseen sopivimpana pitkiä, avoimia haastatteluja. Käytännön syistä tämän tutkimuksen haastattelussa oli kuitenkin oltava rakennetta. Haastateltavat olivat kiireisiä terveydenhuollon asiantuntijoita, ja rajallisessa ajassa oli kerättävä paljon tietoa. Strukturoimaton haastattelu olisi ollut haastateltaville liian aikaa vievä (Gill ym. 2008). Puolistrukturoitu haastattelu on joustava ja monipuolinen tiedonkeruumenetelmä, jota voidaan käyttää tutkimusaineiston hankkimiseen tiettyihin, tarkkoihin tutkimuskysymyksiin vastaamiseksi, ja tästä huolimatta mahdollistaa se, että haastateltavat voivat ilmaista itseään vapaasti (Kallio ym., 2016).

Puolistrukturoiduin asiantuntijahaastatteluin voitiin varmistaa, että tutkimusmateriaali keskittyi tiedon laatuun. Tästä huolimatta oli mahdollista räätälöidä kysymyksiä sopimaan paremmin tietyille haastateltaville ja heidän kokemuksiinsa (Cridland ym., 2015; Kallion ym., 2016 mukaan). Asiantuntijahaastattelu ei ole oma menetelmänsä, mutta siinä on omat erityishaasteensa (Alastalo ym., 2017.) Asiantuntija voidaan määritellä henkilöksi, joka hallitsee oman alansa tietämyksen niin hyvin, että pystyy ratkaisemaan ongelmia, mutta myös *tunnistamaan ongelmien syitä* ja ongelmanratkaisun periaatteita tällä alalla sekä korjaamaan niitä (Pfadenhauer, 2009).

Asiantuntijahaastatteluun valmistautuminen ja sen toteuttaminen poikkeaa maallikoiden haastattelemisesta, sillä tutkijan on päästävä käsiksi asiantuntijatietoon. Pääsyn edessä on erilaisia esteitä kuin maallikoita haastateltaessa. Tärkeää on ensinnäkin kentälle pääsy, se, että saa kontaktin oikeisiin ihmisiin, jotka voivat avata pääsyn asiantuntijoiden luo. Toiseksi tärkeää on hankkia erinomaiset tiedot aiheesta ja pyrkiä näennäisasiantuntijan rooliin. Näin tavoitellaan pääsyä tietoon (episteeminen pääsy). (Alastalo ym., 2017; Pfadenhauer, 2009; Bogner, Littig & Menz, 2009.) Riskinä on jääminen ”asiantuntijamuurin” taa (Alastalo ym., 2017).

4.3 Tiedonkeruun toteutus

Seuraavassa kuvataan tiedonkeruun toteutusta suunnittelu- ja esitietojen keruuvaiheesta valmiiseen aineistoon asti.

4.3.1 Esitietojen ja dokumentaation keruu

Tutkimuksen suunnittelun pohjana olleesta aineistosta käytetään tässä sanaa esitieto. Se kuvaa tutkijan ymmärrystä ja tulkintaviitekehystä. Tutkimuksen esitiedot koostuivat

aiemmasta tieteellisestä kirjallisuudesta, julkisista dokumenteista ja keskusteluista avainhenkilöiden kanssa.

Aiempi kirjallisuus auttaa muotoilemaan hyviä, riittävän tarkkoja tapaustutkimuksen tutkimuskysymyksiä ja rakentamaan teoriaa vahvemmalle perustalle (Yin, 2018). Se auttaa keskittymään tiedonkeruun suunnittelussa olemassa olevan kirjallisuuden aukkoihin (Barriball & While, 1994), ja sillä on tärkeä merkitys puolistrukturoitujen haastattelujen rungon suunnittelussa (Kallio ym., 2016). Muu esitieto palvelee valmistautumisessa asiantuntijahaastatteluihin. Keräämällä etukäteen mahdollisimman runsaasti esimerkiksi haastateltaviin, organisaatioon ja organisaation dokumentteihin liittyvää faktuaalista dokumentaatiota ja muuta materiaalia varmistetaan se, että haastatteluissa voidaan keskittyä vain tarvittavaan sisältöön. (Alestalo & Åkerman, 2010.)

Dokumenttiaineistoa kerättiin internetistä kesäkuusta 2019 vuoden 2020 syksyyn. Dokumenttiaineisto palveli tässä tutkimuksessa paitsi esitietona myös tutkimusmateriaalina haastattelujen rinnalla. Se koostui internetissä julkisesti saatavilla olevista dokumenteista (taulukko 7). Urologian tietoallashankkeen perustamisesta, seurannasta ja tuloksista ei ollut saatavilla virallista dokumentaatiota tutkijan käyttöön, mikä rajoitti kokonaiskuvan saamista hankkeesta ja sen kontekstista ennen haastatteluvaihetta.

Dokumentteihin perehtymisen lisäksi tutkija kävi ennen haastatteluvaihetta keskusteluja avainhenkilöidensä kanssa puhelimessa kesäkuussa 2019 (tilastotieteilijä I4, ks. taulukko 9) ja yhteisessä tapaamisessa syyskuussa 2019 (tietoarkkitehti I3 ja tilastotieteilijä I4, ks. taulukko 9). Kesäkuun puhelinkeskustelussa tutkija sai tietoa urologian hankkeesta ja sen lähtökohdista, aikajananäkymän kehittämisestä ja tietolähteistä. Syyskuun tapaamisessa keskusteltiin hankkeesta lisää, ja tutkija pystyi tarkentamaan dokumentaation perusteella saamaansa kuvaa urologian tietoallashankkeesta ja tietopalvelun prosesseista. Näiden keskustelujen lisäksi tietoja tarkennettiin sähköpostitse raportointivaiheessa avainhenkilöiltä, haastatelluilta ja 2M-IT:ltä.

4.3.2 Kysymysrungon suunnittelu

Puolistrukturoidun haastattelurungon suunnitteluun ei ole yhtä oikeaa tapaa. Yleensä haastattelu sisältää pääteemoja ja jatkokysymyksiä. Pääteemat kattavat tutkimuksen aiheen pääsisällön ja mahdollistavat sen, että haastateltava voi vapaasti jakaa kokemuksiaan ja käsityksiään. Jatkokysymyksiä käytetään moniin eri tarkoituksiin yleisten pääteemojen ohella. (Kallio ym., 2016.) Kallion ym. (2016) mukaan haastattelurungon suunnitteleminen viisiportaisen mallin mukaan parantaa laadullisen tutkimuksen luotettavuutta, ja näin suunniteltu haastattelurunko on käyttövarma.

Tämän tutkimuksen haastattelurungon suunnitteluprosessi kuvataan taulukossa 8. Haastattelurunko (liite 2) perustui kahdeksaan avoimeen kysymykseen, joiden avulla selvitettiin haastateltavan kokemusta ja tietoa urologian tietoallashankkeesta ja tiedon käytöstä, tiedon laatuvaatimuksia, tiedon jalostusprosessia ja haastateltavan tuntemia prosessin osa-alueita sekä tiedon hyödyntämisen ongelmia ja niiden syitä.

Kallion ym. (2016) ideaalimallin noudattaminen osoittautui tutkimuksen kontekstissa haastavaksi. Mahdollisten haastateltavien lista oli varsin lyhyt, joten yhtäkään heistä ei voinut ”tuhlata” varsinaiseen pilottiin. Erillinen pilotti olisi ollut erityisen hyödyllinen, koska tutkijan saatavilla oleva esitieto organisaatiosta itse hankkeen ja haastateltavien työn käytäntöjen osalta ei ollut kovin yksityiskohtaista. Haastatteluvaiheen tehtäväksi jäi

näin myös kerätä faktuaalista informaatiota, ja toimia dokumentaation pohjana. Haastattelurungon validointi pilotoimalla korvattiin siten, että se luetettiin aihealueen asiantuntijalla. Lisäksi ensimmäiset haastattelut toimivat osittain pilotin tavoin. Ensin haastateltiin asiantuntijalääkäriä ja IT-asiantuntijaa, joilla oli oletettavasti vähän viimeaikaista käytännön kokemusta itse urologian hankkeesta, jotta tutkija olisi paremmin valmistautunut tärkeimpien informanttien haastatteluihin.

TAULUKKO 7 Dokumenttiaineisto

Lähde	**Tyyppi**	**Keskeinen sisältö**
Kortekangas (2016)	Esitelmämateriaali	TYKS:in tietoallaskonsepti
Virkki (2017)	Esitelmämateriaali	VSSHP:n tietoaltaan lähdetiedot, urologian potilaan aikajana
Valo, Juhana (2018)	Esitelmämateriaali	2M-IT:n tietoallas
Hammais, Varjonen & Virkki (2018)	Dokumentaatioraportti	Auria Tietopalvelun prosessit ja datalähteet
Darst, Hakala & Kaski (2018)	Arviointiraportti	Isaacus-hankkeen tietoallasratkaisut
Ettala ym. (2018)	Käsikirja	Urologisten sairauksien hoidon laadun mittaaminen
Sitra (2018)	Verkkosivu	Tietoa Isaacus-esituotantohankkeista
Sitra (2019)	Pilottihankkeen loppuraportti + liite	VSSHP:n tietoaltaan tiedon hyödyntäminen kustannusvaikuttavuuden mittaamiseen kansallisin mittarein (KUVA-mittaristo)
Auria Tietopalvelu (2020a ja 2020b)	Verkkosivut	VSSHP:n tietoaltaan aineistot ja niiden sisältö, tietopalvelun organisaatio ja tehtävät

Barribalin ja Whilen (1994) mukaan pilottihaastattelujen perusteella on arvioitava kysymysten toimivuutta ja haastattelijan toimintaa. Ensimmäisissä haastatteluissa ilmeni, että alkuperäinen nimi, jolla hanke tutkijalle oli esitelty, ”urologian hoitopolku”, oli harhaanjohtava, sillä hoitopolku on mahdollista ymmärtää koko hankkeen sijaan kapeammin vain pilotoitavaksi sovellukseksi, näkymäksi, joka on urologien käytössä tällä hetkellä. Potilaan näkökulmasta hoitopolku ei myöskään ole oikea ilmaisu, sillä se ei tässä tapauksessa sisällä perusterveydenhuollon tietoja. Lisäksi haastateltavat eivät tunteneet juurikaan käytännössä urologian tietoallashanketta. Kävi myös ilmi, että jotkin jatkokysymysten rajaukset eivät haastattelussa lainkaan toimineet.

Kokemukset ensimmäisistä haastatteluista olivat muutoin pääosin hyvät: haastateltavat kertoivat kokemuksiaan tiedon laadusta ja laatuongelmista tuoden esiin erilaisia näkökulmia. Myös organisaation sisäistä kriittistä ajattelua tuotiin esiin. Ensimmäisten haastattelujen tunnelma oli rauhallinen, haastattelija antoi tarpeeksi tilaa haastateltavan puheelle, ja aiheiden kehittelylle sekä asioiden mieleen palauttamiselle. Haastattelija valitsi tai hänelle suotiin haastattelussa valistuneen maallikon asema, jolloin aiheissa päästiin

osin syvemmälle ja tarkemmalle tasolle, tosin yksityiskohtaisemmasta teknisestä tiedosta olisi mahdollisesti ollut hyötyä.

TAULUKKO 8 Haastattelurungon suunnitteluprosessi Kallion ym. (2016) viisiportaisen mallin mukaan

Suunnitteluprosessin vaihe	**Vaiheen toteutus**
1. Sen arvioiminen, onko puolistrukturoitu haastattelu tutkimuskysymysten kannalta perusteltu menetelmävalinta	Luvussa 4.2 on arvioitu menetelmävalintaa suhteessa tutkimuskysymyksiin sekä tutkijan ja tutkimuksen kohteena olevan organisaation resurssien näkökulmasta.
2. Aiemman tiedon kerääminen ja hyödyntäminen	Kirjallisuuskatsaus kokoaa tutkijan esitiedon tieteellisestä tutkimuksesta (luvut 2 ja 3) Tutkija keräsi tietoa VSSHP:n tietoaltaasta, prosesseista ja tutkimuskontekstista lukemalla kaiken avoimesti internetissä saatavilla olevan tiedon ja keskustelemalla tutkimuksen yhteyshenkilöiden kanssa organisaatiossa. (kappale 4.3.1) Itse urologian hankkeesta ei ollut ennen tiedonkeruuta saatavilla kirjallista dokumentaatiota tutkijan käyttöön.
3. Alustavan haastattelurungon muotoilu	Alustava haastattelurunko muotoiltiin siten, että se etenee aihealueittain kevyemmästä ja helpommasta abstraktimpaan ja monimutkaisempaan.
4. Alustavan haastattelurungon pilotointi	Ei varsinaista pilotointia. Kysymyksiä ja niiden ymmärtämistä testattiin ulkopuolisella data-analyytikon työtä ja lääketieteellistä tutkimusta tehneellä henkilöllä. Pilotin puuttuminen huomioitiin haastattelujen järjestyksessä. Haastattelurunkoa sovellettiin ja muokattiin haastateltavan tehtävien ja kokemusten mukaan sekä ensimmäisistä haastatteluista saatujen kokemusten perusteella.
5. Lopullisen haastattelurungon esittely	Haastattelurunko on esitelty liitteessä 2.

Lopullinen haastattelurunko esitellään liitteessä 2. Haastattelun alussa määriteltiin haastattelun kohteena olevan hankkeen rajaus ja keskusteltiin nimestä ”urologian hoitopolku”, jota oli käytetty haastateltavien rekrytoinnissa. Jatkokysymykset vaihtelivat haastattelukohtaisesti.

4.3.3 Haastateltavien rekrytointi, informointi ja motivointi

Haastateltavien valinta vaikuttaa suuresti kerätyn tiedon ja faktojen laatuun asiantuntijahaastatteluissa. Ohjenuorana valinnalle on pidettävä asiantuntijoiden tietämystä ja kokemusta. (Alestalo & Åkerman, 2010.) Täten tutkimuksessa tärkeimpiä haastateltavia olivat ne asiantuntijat, jotka tunsivat urologian klinikan hankkeen ja tiedon parhaiten, ja joilla oli riittävästi kokemusta tuon tiedon käytöstä. Nämä tiedon avainkäyttäjät ja -tuottajat olivat

kokeneita kliinikoita ja IT-asiantuntijoita. Organisaation sisäisillä tutkimuksen yhteyshenkilöillä ja portinvartijoilla oli suuri vaikutus haastateltavien rekrytoinnin onnistumiseen. Asiantuntijahaastattelussa tällaiset asiantuntijat, jotka tasoittavat tietä tutkijalle, ovat ratkaisevan tärkeitä (Alestalo et al., 2017). Tutkija sai yhteyshenkilöiltään listan 14 mahdollisesta haastateltavasta, joista kuusi oli kliinikoita ja asiantuntijalääkäreitä, kolme alihankkijan edustajaa ja neljä VSSHP:n IT-asiantuntijaa ja analyytikkoa sekä yksi taloushallinnon asiantuntija. Näistä henkilöistä haastateltiin lopulta kahdeksan, joista puolet lääkäreitä ja puolet IT-asiantuntijoita. Haastateltavat on esitelty taulukossa 9.

Mahdollisia haastateltavia lähestyttiin ensin alustavasti sähköpostitse kuvaillen lyhyesti haastattelun aihealuetta ja viitaten yhteystietojen saamiseen haastateltavan tuntemalta organisaatiossa toimivalta henkilöltä. Haastattelujen sopiminen tehtiin pääosin sähköpostitse, ja halukkaita oli useita. Heille lähetettiin ennen haastattelua tiedote tutkimuksesta sekä tietosuojaseloste. Portinvartijat, jotka mahdollistivat tutkijan sisäänpääsyn organisaatioon, olivat tärkeässä asemassa tutkimusluvan myöntämisen ja haastateltavien osallistumismotivaation kannalta. Yksin heidän osallistumisensa tutkimukseen antoi kuvan siitä, että se oli vaivan arvoista. Tutkimuskysymysten tärkeys organisaatiolle oli motivoiva tekijä haastateltaville, koska he voivat hyötyä tutkimuksen tuloksista kehittäessään omaa työtään ja sairaalan käytäntöjä. Lisäksi tutkijan itsensä antamalla vaikutelmalla on suuri vaikutus suostumiseen haastateltavaksi ja keskusteluhalukkuuteen haastattelun aikana (Alestalo et al., 2017).

Haastateltujen lopulliseen määrään vaikutti paitsi tutkimusasetelma, myös haastatteluun saatavilla olevien asiantuntijoiden lukumäärä ja osallisten rajalliset resurssit. Koska tutkimuksella oli ennen haastatteluja rajattu teoreettinen perusta ja se keskittyi ainoastaan yhteen tapaukseen, haastattelujen määrä voitiin pitää suhteellisen pienenä. Toisaalta haastateltavien toisistaan poikkeava tausta eri alojen asiantuntijoina eri puolilla organisaatiota ennakoi tarvetta hieman useammalle haastattelulle, samoin kuin kokematon ja kontekstia rajallisesti tunteva haastattelija ja haastatteludialogin suhteellisesti heikompi laatu. (Malterud, Siersma, & Guassora, 2016.) Lopulta haastateltiin kaikki haastatteluun halukkaat henkilöt.

4.3.4 Haastatteluaineisto

Tutkija toteutti haastattelut VSSHP:n tutkimusluvalla syys-marraskuussa 2019. Haastattelu pyrittiin tekemään haastateltavan työpaikalla, ja sille oli pyydetty varaamaan aikaa tunti. Haastatteluaineisto muodostui kahdeksasta noin 40–70 minuutin mittaisesta haastattelusta (taulukko 9). Kuusi haastattelua tehtiin kasvotusten haastateltavien työpaikalla, kaksi haastattelua internetpuhelun kautta.

Haastatteluaineisto oli heterogeeninen, sillä haastateltujen roolit hankkeessa poikkesivat toisistaan. Koska monet haastatelluista eivät tunteneet urologian hanketta kovin hyvin, haastattelujen sisältö koski lähes poikkeuksetta tietoallastiedon hyödyntämisen ongelmia ja niiden syitä yleisemmin tutkimuksessa ja tiedolla johtamisessa. Urologian hankkeessa eri vaiheissa mukana olleiden haastattelujen avulla saatiin kuitenkin aineistoa urologian klinikalla kohdatuista tiedon laatuongelmista ja niihin kehitetyistä konkreettisista ratkaisuista. Haastattelujen jälkeen äänitykset purettiin tekstimuotoon. Haastateltaville annettiin mahdollisuus vaikuttaa siihen, missä muodossa heidän tietonsa taulukossa 9

raportoitiin. Lääkärien koodi alkaa kirjaimella L, IT-asiantuntijoiden kirjaimella I. Haastateltaville annettiin myös mahdollisuus saada haastattelunsa litterointi tarkistettavaksi.

TAULUKKO 9 Haastatteluaineisto

Koodi	**Ammattiryhmä**	**Organisaatio**	**Rooli urologian tietoallashankkeessa**	**Haastattelun tyyppi**	**Haastattelun kesto (minuuttia)**
L1	Tutkimusjohtaja	TYKS	Tärkeä rooli tietoaltaan pystytyksessä, tietoallastyöryhmän puheenjohtaja. Hankkeen alkuun panijoita, mahdollistaja ja tukija.	käynti	53
I2	Data science -kehityspäällikkö	VSSHP	Ei virallista roolia hankkeessa, kerännyt alkuvaiheessa patologian tietoja urologian hankkeeseen	käynti	63
I3	Tietoarkkitehti	VSSHP	Tiedon hahmottaminen keskustelemalla urologien kanssa ja tietokantojen rakentaminen	käynti	60
I4	Tilastotieteilijä	VSSHP	Käytännön kehittäjä ja visioija, käyttöliittymän ja siihen liittyvän analytiikan toteuttaja	videopuhelu	70
L5	Syöpälääkäri, esimiesasemassa	TYKS	Seurannut hankkeesta käytävää keskustelua	käynti	42
L6	Urologian erikoislääkäri	TYKS	Hankkeen vastuuhenkilö urologian klinikalla	käynti	62
I7	Tuotepäällikkö	2M-IT	Ollut palveluntarjoajan edustajana mukana lähdejärjestelmätiedon valmistelussa urologian hankkeen alkuvaiheesta alkaen.	käynti	59
L8	Urologian klinikan ylilääkäri	TYKS	Toimeenpanija ja hankkeen tiedolla johtaja	puhelinhaastattelu	41

4.4 Aineiston analyysi

Tapaustutkimuksen analyysi perustuu analyyttisiin kysymyksiin, joihin vastataan käyttäen kaikkea tutkimusaineistoa (Yin, 2018). Tässä tutkimuksessa pyrittiin hahmottamaan tapauksen säännönmukaisuuksia (pattern matching) ja loogista mallia (logical model) (ks. Yin, 2018). Menetelmänä ja joustavana teoreettisena viitekehyksenä käytettiin laadullista sisällönanalyysiä, joka sopii monenlaisille tutkimusotteille ja -aineistoille (Tuomi & Sarajärvi, 2018). Laadullinen sisällönanalyysi pyrkii tekstiaineiston sisällön subjektiiviseen tulkintaan, jossa systemaattisen luokitteluprosessin keinoin koodataan ja tunnistetaan teemoja ja säännönmukaisuuksia (Hsieh & Shannon, 2005). Hsieh'in & Shannon'in (2005) mukaan aineiston luokittelukoodit ja koodijärjestelmät voidaan johtaa joko suoraan tekstiaineistosta (conventional content analysis) tai teoriasta ja relevanteista tutkimuslöydöksistä (directed content analysis). Lisäksi se voi sisältää tekstin sisällön tai avainsanojen laskemista ja vertailua (summative content analysis). Ensiksi mainittuja voidaan nimittää myös aineisto- ja teorialähtöiseksi sisällönanalyysiksi (esim. Sarajärvi & Tuomi, 2018; Latvala & Vanhanen-Nuutinen, 2001) tai induktiiviseksi ja deduktiiviseksi sisällönanalyysiksi (Elo ym., 2014).

Analyysimenetelminä yhdistettiin aineisto- ja teorialähtöistä sisällönanalyysiä. Eri tutkimuskysymyksiin vastaamiseksi sovellettiin erilaista lähestymistapaa. Tiedon laatuun vaikuttavat seikat (laatuongelmien syyt) koodattiin aineistolähtöisesti, sillä olemassa oleva kirjallisuus on hajanaista ja suppeaa (ks. Hsieh & Shannon, 2005). Analyysiyksikkönä oli tiedon laatuun liittyvä, syy-seuraussuhteita sisältävä asiakokonaisuus, virke tai keskustelunpätkä.

Tiedon laatuominaisuuksien koodaaminen tehtiin teorialähtöisesti. Siinä käytettiin Wangin ja Strongin (1996) tiedon laadun käsiteviitekehystä täydennettynä siitä pois jätetyillä laatuominaisuuksilla jäljitettävyys, tietolähteiden moninaisuus, kustannustehokkuus, teknisen käsittelyn helppous ja joustavuus (ks. taulukko 1), jotka ovat aiemman kirjallisuuden mukaan big datan osalta relevantteja ominaisuuksia (esim. Merino ym. , 2016; Clarke, 2016). Kustannustehokkuuden määritelmää täsmennettiin koskemaan paitsi tiedonkeruun myös kaikkien tiedon käsittelyn vaiheiden kustannustehokkuutta. Näin oli mahdollista selvittää, miten viitekehys sopii kuvaamaan kliinisen tiedon ja big datan laatuominaisuuksia tietoallasympäristössä ja arvioida tarvetta viitekehyksen ulkopuolisille laatuominaisuuksille (ks. Hsieh & Shannon, 2005). Tiedon laatuongelmien syntyprosessien säännönmukaisuuden (data quality problem patterns) kuvaamisessa ja visualisoinnissa hyödynnettiin lisäksi lähtökohtana Strongin ym. (1997) tutkimusta. Tämä auttoi varmistamaan, että analyysi ottaa huomioon tiedon laadun kontekstuaaliset puolet (ks. Hsieh & Shannon, 2005), mikä oli koko tutkimuksen tarkoituksena.

Analyysi alkoi esittämällä aineistolle hankkeesta ja tiedonjalostusprosessista kysymyksiä, joiden vastaukset olivat jääneet varsinaisessa esitietojen keruuvaiheessa puutteellisiksi, minkä jälkeen jatkettiin tiedon laatuongelmien ja niiden syy-seuraussuhteiden säännönmukaisuuksien analyysiin. Aineisto luettiin useita kertoja läpi eri näkökulmista pyrkien löytämään kontekstuaalisen tiedon laatuprosesseja ja niiden kokemuksellisia tulkintoja. Aineiston järjestelyssä erilaisin taulukoin käytettiin Excel-ohjelmaa. Lisäksi säännönmukaisuuksia ja logiikkaa pyrittiin kuvaamaan visuaalisesti syy-seurausketjuina käyttäen yEd Graph Editor -ohjelmaa. Analyysissä huomioitiin haastattelujen luonne asiantun-

tijahaastatteluina. Suhteessa faktojen ja vuorovaikutuksen olemukseen valittiin Alestalon ja Åkermanin (2010) ehdottama keskitien lähestymistapa: faktat ovat ”yhdessä tehtyjä” ja vuorovaikutus on haastatteluaineiston luonteenomainen piirre. Samaan tapaan suhtauduttiin aineistolähtöisyyteen. Tutkijan havainnot eivät aineistolähtöisessä analyysissä ole objektiivisia ja puhtaita, vaan niihin sisältyy teoriapitoista tulkintaa (Sarajärvi & Tuomi, 2018).

4.5 Laadullisen tutkimuksen luotettavuus

Kvantitatiivisen tutkimuksen ihanteena on objektiivisuus ja yleistettävyys, mutta kvalitatiivinen tutkimus on aina subjektiivista ja kontekstisidonnaista. Tämän vuoksi sen luotettavuuden arviointi eroaa kvantitatiivisesta tutkimuksesta. (Whittemore, Chase & Mandle, 2015.) Laadullisen tutkimuksen piirissä on erilaisia tulkintoja siitä, miten sen luotettavuutta arvioidaan (Tuomi & Sarajärvi, 2018). Osa tutkijoista soveltaa kvantitatiivisen tutkimuksen reliabiliteetin (tulosten toistettavuus) ja validiteetin (tilastollisen mallin osuvuus tai selitysvoima) käsitteitä soveltuvin osin, mutta usein näille käsitteille annetaan uusi sisältö (Eskola & Suoranta, 1999; Mäkelä, 1990; Hirsjärvi & Hurme, 2000). Tällöin laadullisen tutkimuksen reliabiliteetti riippuu analyysin systemaattisuudesta, validiteetti tutkijan tekemien tulkintojen uskottavuudesta (Ruusuvuori ym., 2010; Hirsjärvi & Hurme, 2000). Validiteettia ja reliabiliteettia on kuitenkin myös kritisoitu lähtökohtiensa ja alansa perusteella sopimattomiksi laadulliseen tutkimukseen (Tuomi & Sarajärvi, 2018).

Laadullisen tutkimuksen omia luotettavuuden arviointikriteereitä ovat uskottavuus (credibility) – miten hyvin tutkijan konstruktiot vastaavat tutkittavien todellisuutta, siirrettävyys (transferability) – miten hyvin tulokset ovat siirrättävissä toiseen kontekstiin, varmuus (dependability) – miten hyvin tutkimuksessa on otettu huomioon erilaiset tutkimukseen vaikuttavat ja ennakoimattomat tekijät, ja vahvistuvuus (confirmability) – saako tutkimus tukea ulkopuolisen arvioijan tai muun tutkimuksen tekemistä tulkinnoista. Nämäkin ovat vain yksi tapa määritellä ja suomentaa kriteerit. (Tuomi & Sarajärvi, 2018; Eskola & Suoranta, 1999.)

Laadullisen tutkimuksen tulosten uskottavuutta voidaan vahvistaa triangulaation avulla eli tutkimalla samaa ilmiötä eri näkökulmista (Patton, 2002; Eskola & Suoranta, 1999). Tapaustutkimuksessa triangulaatio on keskeinen tulosten vahvistamisen tapa (Yin, 2018). Tässä tutkimuksessa käytettiin kahdenlaista aineistoa, haastatteluja ja dokumentteja. Eri analyysimenetelmät ja tavat järjestellä ja analysoida aineistoa (luokittelu, visualisointi) auttoivat katsomaan aineistoa ja tuloksia uusin tavoin. Myös eri rooleissa eri puolilla tapausorganisaatiota toimivat haastateltavat tarjosivat toisistaan poikkeavia näkökulmia (ks. Myers & Newman, 2007). Haastateltavilla oli lisäksi mahdollisuus korjata oman haastattelunsa mahdollisia väärinymmärryksiä ja virheellisiä faktoja litteroidusta tekstistä (ks. Latvala & Vanhanen-Nuutinen, 2001; Yin, 2018). Siirrettävyyteen pyrittiin riittävän tarkalla kontekstin kuvaamisella, varmuuteen puolestaan esitietojen keruulla ja haastattelurungon riittävän joustavalla muotoilulla.

Vaikka laadullisen tutkimuksen luotettavuuden arvioinnin tavat ja käsitteet vaihtelevat paljon, laadullista tutkimusta voidaan arvioida ennen kaikkea *tutkimusprosessin* luotettavuuden näkökulmasta. Koska tutkija on laadullisen tutkimuksen tärkein tutkimusväline, arvioinnin kohteena ovat paitsi aineistonkeruu- ja analyysimetodien ja tekniikoiden

käyttö myös tutkijan uskottavuus ja pätevyys (Patton, 2002; Eskola & Suoranta, 1999). Tutkimusprosessin arvioimiseksi on pyritty antamaan riittävän tarkka kuvaus tutkimuksen kulusta eri vaiheineen sekä tehtyjen valintojen perustelut (ks. Hirsjärvi & Hurme, 2000; Eskola & Suoranta, 1999).

4.6 Urologian tietoallashanke

Tässä alaluvussa kuvataan tapaushanketta dokumenttien ja haastattelujen valossa. Virallisen kuvauksen puuttuessa hanketta nimitetään tässä urologian tietoallashankkeeksi, mikä sisällyttää siihen ne erilaiset tavat, joilla tietoaltaan tietoa on urologian klinikalla hyödynnetty hankkeen kontekstissa. Se myös sisällyttää toimijoiden hankkeelle antamat erilaiset merkitykset. Koska tutkijan esitieto oli puutteellista ja ymmärrys tutkimuksen kohteesta muotoutui sitä tehdessä, suuri osa näistä tiedoista on kerätty haastattelujen yhteydessä.

Hanke esiteltiin tutkijalle alun perin nimellä ”Urologian hoitopolku”. Hankkeessa oli kyse pääasiassa VSSHP:n oman tietoaltaan tietojen hyödyntämisestä urologian klinikan toiminnassa. Yleisluotoisen kuvan hankkeen alkuvaiheesta ja VSSHP:n tietoaltaasta tutkija sai Virkin (2017) esitelmämateriaalista ja VSSHP:n kliinisen tietopalvelun tiedonjalostusprosessin kuvauksesta (Hammais, Varjonen & Virkki, 2018). Tätä täydensivät keskustelut Auria tietopalvelujen yhteyshenkilöiden kanssa (I3, I4). Hankkeen kirjallinen dokumentaatio rajoittui haastateltavien mukaan IT-palveluntarjoajan (Medbit Oy/2M-IT) kanssa tehtyihin alkuperäisiin sopimuksiin ja vaatimusmäärittelyihin, joita tutkijalla ei ollut käytössään.

4.6.1 Urologian tietoallashankkeen tausta ja eteneminen

Urologian tietoallashankkeen taustalla on ollut kaksi erillistä potilastiedon laatuun liittyvää epäkohtaa. Ensinnäkin kokonaiskuvan luominen potilaan tilanteesta eri potilastietojärjestelmien pohjalta on aikaa vievää. Suuri osa potilastiedoista on rakenteettomassa muodossa, erityisesti potilaskertomustekstinä, jolloin tarvittavan tiedon löytäminen on hankalaa ja hidasta. Toiseksi hoitoon tarkoitetut potilastietojärjestelmät eivät mahdollista tiedon keräämistä hoidon laadun seuraamiseksi, tai kerätyn tiedon laatu on huono. Olennaista tietoa puuttuu tai se ei ole luotettavaa ja oikeellista. Potilastietojärjestelmien tiedon ei siis koeta olevan tarpeeksi nopeasti ja tiiviissä muodossa silmäiltävissä (tiedon saatavuuden ja representationaalisen laadun ongelmat), ja siitä puuttuu tärkeää tietosisältöä hoidon laadun arvioimiseksi tai tällaisen tiedon laatu on huono (tiedon sisäisen ja kontekstuaalisen laadun ongelmat).

Aluksi hankkeen keskiössä oli tiedon jalostaminen potilaan hoidon ja johtamisen tueksi, mutta myöhemmin hanke laajeni eturauhassyövän ja muiden urologisten syöpien hoidon laaturekisterihankkeeksi. Isaacus-hankkeen aikana tehtiin vaatimusmäärittely niin sanotulle urologian hoitojanaesimerkille (aikajananäkymä). Sen avulla oli tarkoitus pilotoida tietoaltaan tietoa hyödyntävää erikoissairaanhoidon näkymää, joka kerää yhteen eri

potilastietojärjestelmien tiedot tiivistetysti visuaaliselle aikajanalle potilaskohtaisesti tai aggregoidusti. Oman laaturekisterin tarve syntyi, kun kokeilun jälkeen päätettiin luopua kalliiksi ja tietosisällöltään omiin tarpeisiin sopimattomaksi koetusta kaupallisesta, BCB Medicalin laatusovelluksesta. Suomen Urologiyhdistys oli perustanut jo vuonna 2017 oman laatutyöryhmän pohtimaan urologisten sairauksien hoidon laatua ja sen mittaamista (ks. Ettala ym., 2018). VSSHP:ssa oli urologian tietoallashankkeen pohjalta olemassa hyvät edellytykset kansallisen eturauhassyövän laaturekisterin prototyypin rakentamiselle. Kehitysvaiheessa oleva urologian aikajanasovellus, tietoaltaan tietojen jalostamien ja yhteistyö tietopalvelun kanssa muodosti pohjan kehitystyölle. Auria Tietopalvelu alkoi kehittää laaturekisterisovellusta, ja se valmistui pääosin vuoden 2018 aikana.

VSSHP:n urologian klinikan laaturekisterihankkeella on läheinen yhteys vuonna 2018 käynnistettyyn THL:n eturauhassyövän laaturekisterihankkeeseen, jonka vetäjänä VSSHP:n urologian tietoallasankkeen vetäjä työskentelee osa-aikaisesti. Kansallinen hoidon laadun arvioimisen näkökulma onkin painottunut hankkeessa yhä enemmän sen edetessä. Hankkeen avulla kehitetään työkaluja kansallisen vertailutiedon keräämiseksi eturauhassyövän hoidon laadusta, jotta eri yksiköitä ja hoitoja voidaan arvottaa ja vertailla.

Aikajananäkymän tietosisältöä ja visuaalisuutta kehitettiin noin puolentoista vuoden ajan vuosina 2017–2019 urologian klinikan lääkärien ja Auria-tietopalvelun asiantuntijoiden yhteistyönä prototyyppien sekä kliinikoiden näkemyksen ja palautteen pohjalta. Kliinikoiden palautteen pohjalta on tehty lisäyksiä, tarkennuksia ja muutoksia. Toimiva tekninen ratkaisu löytyi muutaman vaihtoehdon kokeilemisen jälkeen. Näkymä on virallisesti otettu käyttöön urologian poliklinikalla 16.9.2019.

Näkymän kehittämisestä hanke eteni vuonna 2018 yhä enemmän hoidon laadun mittaamisen ja tiedonkeruun suuntaan. Kun aluksi keskipisteenä oli yksilönäkymä, nyt keskeistä oli yksikön tilastointi ja raportointi, jota näkymä tukee. Laaturekisterisovellukseen viedään tietoa useilla potilaskohtaisilla lomakkeilla, joissa kerätään tietoa hoitosuunnitelmasta, käytännön hoidosta, kotiutuksesta ja seurannasta. Kerättyjen tietojen avulla on tarkoitus raportoida eturauhassyöpäpotilaiden leikkausmäärät, komplikaatiot ja infektiot kuukausitasolla.

4.6.2 Hankkeen mahdollistajat

Urologian tietoallashanke on vaatinut sitä, että tietoa voidaan laillisesti käyttää, sitä on saatavilla ja osataan hyödyntää. Tärkeitä hankkeen mahdollistajia ovat olleet tiedon saatavuus, toimiva organisaatio ja tekninen infrastruktuuri sekä tietotaito. Tiedon saatavuutta ovat edistäneet toisiolaki, oma tietoallas, pitkät perinteet sähköisen potilastiedon tallentamisessa sekä laadukas tiedonhallinta. Tiedon prosessoinnin, jalostamisen ja käytön tarpeisiin on olemassa välineet. Tietoaltaan ja muun teknisen infrastruktuurin lisäksi VSSHP on hankkinut tehokkaita grafiikkaprosessoreita muun muassa tekstinlouhintaan. Niitä on alettu käyttää myös kuvien analysoimiseen. Koko arkkitehtuuri on suunniteltu hoidon laadun seurannan näkökulmasta.

Tiedon ja teknologioiden lisäksi on tarvittu ihmisiä ja yhteistyötä. Hankkeen käynnistymisessä keskeisiä toimijoita ovat olleet aktiiviset kliinikot ja sairaanhoitopiirin oma tietopalvelu (Auria Tietopalvelu), jossa on osaamista lääketieteestä ja tietojärjestelmistä ja tunnetaan potilastietojen tallennus- ja käyttämisprosessi. Tietopalvelun ja urologian klini-

kan fyysinen läheisyys TYKS:ssa mahdollistaa sujuvan viestinnän ja yhteistyön. Se on ollut edellytys IT-ammattilaisten ja kliinikoiden yhteisen kielen löytämiselle. Hankkeen onnistuminen on edellyttänyt, että IT-ammattilaisten ja lääkärien on ymmärrettävä toisiaan ja kyettävä läheiseen yhteistyöhön. Erilaisten resurssien tehokas hyödyntäminen organisaation toiminnassa onkin ollut keskeinen hankkeen mahdollistaja.

4.6.3 Hankkeessa tavoitellut hyödyt

Haastateltavat odottavat tietoallashankkeesta hyötyjä monella tasolla, yksittäisille potilaille, lääkäreille, organisaatiolle ja yhteiskunnallisesti. Se on koettu urologian klinikalla huimana parannuksena tiedon laatuun. Potilastietoa voidaan rakenteistaa, ja sen käytettävyys on parempi kuin ennen. Näkymän ansiosta lääkärillä menee pitkään sairastaneen potilaan tietojen selailuun viidestä kymmeneen minuuttia, kun ennen siihen saattoi kulua 20–25. Tärkeät tiedot on helpompi huomata näkymästä kuin selaamalla potilastietojärjestelmiä. Toivotaan, että tulevaisuuden potilastietojärjestelmissä voisi olla samantyyppinen käyttöliittymä.

Toiminnan seuraaminen on parantunut, kun siitä saadaan systemaattista tietoa. Tarkoituksena on luoda raportteja eri tasoille: sairaalan johdolle, klinikan ylilääkärille ja sotepäätökseen. Laaturekisteritietoa on kertynyt vasta vähän, mutta suuntaa antavia tuloksia on jo saatu eturauhassyöpäpotilaiden imusolmukkeiden poistoon liittyvistä komplikaatioista. Oman laaturekisterin etuna on, että sitä voidaan räätälöidä juuri oman klinikan toiminnan tarpeisiin. Laaturekisterin tietosisältö on myös päivittyvä ja mukautuu hoitokäytäntöjen muutoksiin.

Taloudellista hyötyä syntyy lyhyelläkin tähtäimellä, jos näkymän avulla vältetään turhia uusia kokeita ja tutkimuksia, kun tieto esim. aiemmista laboratoriotuloksista on paremmin esillä. Pidemmällä tähtäimellä voidaan kohdistaa yksikön resursseja tehokkaammin. Yksittäisen lääkärin saama palaute antamansa hoidon laadusta on hankkeen tärkein hyöty. Lääkäri voi tiedon avulla nähdä omat virheensä ja oppia niistä. Hanke lisää lääkärien motivaatiota kirjata potilastietoja paremmin, kun he näkevät sen hyötyjä, mikä parantaa jatkossa potilastiedon laatua.

Hankkeen ansiosta eturauhassyövän hoidon laadun kehittäminen mahdollistuu. Hankkeeseen liittyy laajempia hyötyjä, sillä sen avulla pyritään saamaan THL:n eturauhassyövän laaturekisteripilotin mukainen tietokanta CSC:n palvelimelle sekä helppo käyttöliittymä kansalliseen käyttöön. Hankkeessa käytetään avointa lähdekoodia, jotta implementoiminen onnistuisi mahdollisimman pienin kustannuksin. Kun eri yksiköissä hoidon laatua mitataan samoin kriteerein, voidaan vertailla yksiköitä keskenään. Se mahdollistaa alueellisen yhdenvertaisuuden, resurssien ja kustannusten kohdentamisen oikeudenmukaisemmin.

4.6.4 Hankkeen haasteet

Hankkeen haasteet liittyvät tietoallastiedon laatuun ja siitä varmistumiseen. Potilastietojärjestelmiä ei ole suunniteltu retrospektiivisen analyysin tarpeisiin, ja tietoaltaan tiedon saaminen käyttöön vaatii monimutkaisia työvaiheita. Tietoaltaan tietoa hyödyntävällä nä-

kymällä ei ole lääkinnällisen lisälaitteen asemaa (CE-merkintää), koska tiedon oikeellisuutta ei voida lain vaatimalla tavalla varmistaa, ja laki kieltää hoitopäätösten tekemisen sen perusteella. Riskinä on myös, että laaturekisteriin halutaan liian paljon tietoja, kun tälläkin hetkellä lääkärit joutuvat kirjaamaan potilastietoja kahteen kertaan (ns. kaksoiskirjaaminen). Kansallisesti haasteena on, miten saadaan koko kansallinen tieto yhteen. Tähän liittyy muun muassa rekisterinpidon-, rahoituksen ja vastuukysymyksiä.

4.6.5 Haastateltavien näkemyksiä hankkeesta

Haastateltavien rooli urologian tietoallashankkeessa ja se, missä vaiheessa he olivat sitä seuranneet tai olleet siinä mukana, vaikutti siihen, miten he mielsivät hankkeen sisällön ja sen tavoitteet. Urologien näkemyksissä painottui laaturekisteri ja tavoite luoda IT-ratkaisu urologisten syöpien hoidon laatutiedon keräämiseen omalla klinikalla ja kansallisesti. Käytettiin myös nimeä ATP-portaali. Useimmat muut haastateltavat nimesivät hankkeen näkymien kehittämisen pilottihankkeeksi tai prototyypiksi, erikoissairaanhoidon hoito-osuuden näkymäksi, hoitojanaesimerkiksi ja aikajanaksi. Hankkeessa työskentelevät IT-asiantuntijat puhuivat urologien lailla laaturekisterihankkeesta sekä isommasta kokonaisuudesta sen taustalla.

4.7 Tiedonjalostusprosessi

Tiedon laatuun vaikuttavien prosessien ymmärtämiseksi tutkijan oli selvitettävä tiedon synty- ja jalostusprosessin vaiheet ja se, ketkä tietoa missäkin vaiheessa käsittelevät. Jokaisen haastateltavan oma ymmärrys ja kertoma tästä prosessista oli tutkijalle tärkeää haastattelun tulkintaan vaikuttavaa tietoa. Tätä varten kahta lääkäriä ja kaikkia neljää IT-asiantuntijaa pyydettiin piirtämään kuva niistä vaiheista, joissa tieto kulkee, kun sitä halutaan käyttää urologian hankkeessa ja kuvailemaan sanallisesti, mitä eri vaiheissa, varsinkin niissä, jotka oman työn kautta olivat heille tuttuja, tapahtuu. IT-asiantuntijoille kuvan piirtäminen oli helppoa, mutta kumpikaan lääkäri ei sitä tehnyt, koska ei kokenut sitä itselleen luontevaksi, vaan kuvasi vaiheita sanallisesti.

Tässä luvussa on yleiskuvaus hankkeen tiedonjalostusprosessista dokumenttien ja haastattelujen pohjalta koostettuna. Prosessi alkaa tiedon tuotannosta, sisältää tiedon tallennuksen ja varastoinnin sekä erilaiset siirto-, lataus- ja muokkausvaiheet ja päättyy tiedon käyttöön.

4.7.1 Potilastietojen kirjaaminen

Potilastietoja kirjataan moniin erikoissairaanhoidon tietojärjestelmiin esimerkiksi laboratoriossa, kuvantamisen yksikössä, vastaanotolla ja osastolla. Näitä tuotantojärjestelmiä ylläpitävät monet eri järjestelmäntoimittajat (ks. taulukko 6). Kirjaajia ovat lääkärit, hoitajat ja

joissain tapauksissa potilas itse. Eturauhassyövän hoidossa TYKS:issä potilas täyttää itse prostatasyöpäspesifisen potilaan yleisen voinnin esitietojen kyselyn.

Kirjaaminen tietojärjestelmiin lääkärin vastaanotolla tapahtuu seuraavasti: lääkäri kirjoittaa itse tai yleensä sanelee lyhyen decursuksen, diagnoosin, diagnoosinumerot, toimenpiteet ja toimenpidenumerot, jotka osastosihteeri kirjoittaa tekstiksi. Sihteeriltä teksti tulee vielä lääkärille tarkistettavaksi. Saneluun on mahdollista vaihtoehtoisesti käyttää digitaalista puheentunnistusta, jolloin teksti on heti lääkärin tarkistettavissa.

Urologian klinikalla lääkäri täyttää potilastietojärjestelmien tietojen lisäksi potilaskohtaisesti oman toiminnan seurannan raportointilomakkeet (hoitosuunnitelma, käytännön hoito, kotiutus, seuranta) urologian hankkeessa kehitettyyn käyttöliittymään. Siihen täytetään urologian laatuhankkeen kannalta relevantit tiedot, jotka ovat periaatteessa samoja tietoja kuin potilastietojärjestelmissä on, mutta rakenteisessa muodossa. Lääkäri voi pitää vastaanotolla rinnakkain toisella ruudulla potilastietojärjestelmää, laboratorio-ohjelmaa ja kuvantamisohjemaa ja toisella urologian hankkeessa kehitettyä näkymää, jolloin potilaan hoidon vaihetta ja kokonaiskuvaa voi hahmottaa urologian näkymästä ja sen tietoja varmentaa potilastietojärjestelmistä. Näkymästä pääsee ikoneita klikkaamalla katsomaan alkuperäistä potilaskertomustekstiä. Potilaan haastattelun ja laboratoriotutkimusten jälkeen kirjataan lomakkeelle rakenteisesti, missä tilanteessa potilas on.

4.7.2 Tietojen tallennus, varastointi, jalostus ja toissijainen käyttö

Kopio tietojärjestelmien tiedosta siirretään yön yli 2M-IT:lle tietoaltaaseen. Kopio voi olla suora kopio tai jalostettu kopio jonkinlaiseen raportointirakenteeseen. Se toimii raakadatavarantona, jossa tieto on mahdollisimman alhaisella alkiotasolla juuri sellaisena, kuin se on tallennettu potilastietojärjestelmään. Sen vähimmäissäilytysaika on 12 vuotta. 2M-IT:n tehtävänä on pitää alkuperäinen sensitiivinen tieto koskemattomana ja eheänä.

Urologian hankkeessa tietoaltaasta ammennetaan eri potilastietojärjestelmien tietoa tietopalvelujen erilliseen analytiikkaympäristöön mutta osa potilastietojärjestelmien tiedoista siirretään kuitenkin sinne suoraan lähdejärjestelmistä, niin, että se ei käy tietoaltaassa. Tästä vastaa tietopalvelutiimi. Tässä vaiheessa tiedon jalostaminen alkaa. Sitä voidaan yhdistellä eri lähteistä, muokata ja jalostaa eri asteiseksi. VSSHP:ssä tiedon yhdistäminen, puhdistaminen ja yhteismitallistaminen (harmonisointi) ja muu jalostaminen tapahtuu rakenteisessa tietovarastossa, joka on toteutettu avoimella lähdekoodilla (PostgreSQL).

Tietopalvelutiimi tekee tietoaltaan tiedosta helpommin kyseltävissä ja luettavissa olevia tietokantoja, tietokantatauluja ja näkymiä, joissa tiedon yhdistelyä ja muokkaamista tehdään. Urologian hankkeessa on rakennettu näkymä ja käyttöliittymä. Raportointinäkymän eli urologian aikajanasovelluksen tiedonsiirto on tuotteistettu. Omat ennalta määritellyt raportit ovat omassa näkymässään, ja koko ajan kehitetään uusia seurattavia mittareita. Urologian tiedolle tehdään muunnoksia vielä sovellusnäkymän lukemis- ja esittämisvaiheessa, esimerkiksi PSA-mittauksessa saadut arvot on sovitettu tasoitetulle käyrälle, jota on helpompi lukea.

Tiedon raportoinnista ja käytöstä haastateltavat osasivat kertoa vielä melko vähän. Raportointia tehdään eri johtotasoille: klinikan ja sairaalan johdolle sekä sote-päätökseen. Klinikan johtaja on käyttänyt tietoa resurssien hallintaan. On selvitetty, miten syöpätapausten määrä on kehittynyt vuosittain ja kuinka paljon leikkauksia on tehty. Kun tietoa

kertyy riittävästi, sitä on tarkoitus käyttää resurssien kohdentamiseen ja leikkausmenetelmien valintaan. Urologian laaturekisterihankkeen raportointikanta ja raportoinnin visualisoinnit ovat vielä kehittämisvaiheessa. Tarkoituksena on tarjota raportointia omien leikkausten onnistumisesta ja laatuparametreista. Tietoa on tarkoitus käyttää myös tutkimukseen tutkimuslupien puitteissa.

5 TULOKSET: KLIINISEN TIETOALLASTIEDON LAATUONGELMAT OVAT MONINAISIA

Tässä luvussa käydään läpi analyysin tulokset tiedon laatuongelmien syiden mukaan jäsenneltynä. Syyt luokiteltiin aineistolähtöisesti kahteen yläluokkaan sen mukaan, liittyvätkö ne tiedon tallentamiseen potilastietojärjestelmiin vai tiedon varastointi- ja jalostusvaiheeseen. Luokitus on esitelty kuviossa 10. Molemmissa kontekstina on tiedon toissijainen käyttö tutkimukseen ja tiedolla johtamiseen. Laatuongelmat molemmissa käyttötarkoituksissa on mahdollista esittää samassa yhteydessä, sillä tiedon laatuvaatimukset ja laatuongelmat eivät haastateltavien mukaan niissä eroa toisistaan.

Laatuongelmien luokitteluun käytettyä viitekehystä (taulukko 1) täydennettiin aineistosta nousseilla kuudella uudella laatuominaisuudella: *rakenteisuus, tarkkuus, vertailtavuus, yhdisteltävyys, yhtenäisyys* (*tiedon rakenteen yhtenäisyys (vs. pirstaleisuus)) ja eheys (tiedon kulku prosessissa muuttumattomana)*. Tutkijan tekemä koodaus on tuotu näkyviin seuraavalla tavalla: Syiden yhteyteen on merkitty lihavoidulla tekstillä ja nuolin tutkijan tekemät tulkinnat syy-seuraussuhteista. Kaikki laatuongelmat eivät välttämättä ole tapauskonteksissa ilmenneitä ongelmia, vaan tarkoituksena on ollut raportoida myös mahdollisia syy-seurausketjuja haastateltujen näkemysten pohjalta. Syy-seuraussuhteet on koottu kokonaiskuviksi prosessikuvioihin (kuvio 5, tallennusvaihe ja kuvio 6, tiedon varastointi-, jalostus- ja käyttö).

5.1 Potilastiedon tallennukseen liittyvät tiedon laatuongelmat ja niiden syyt

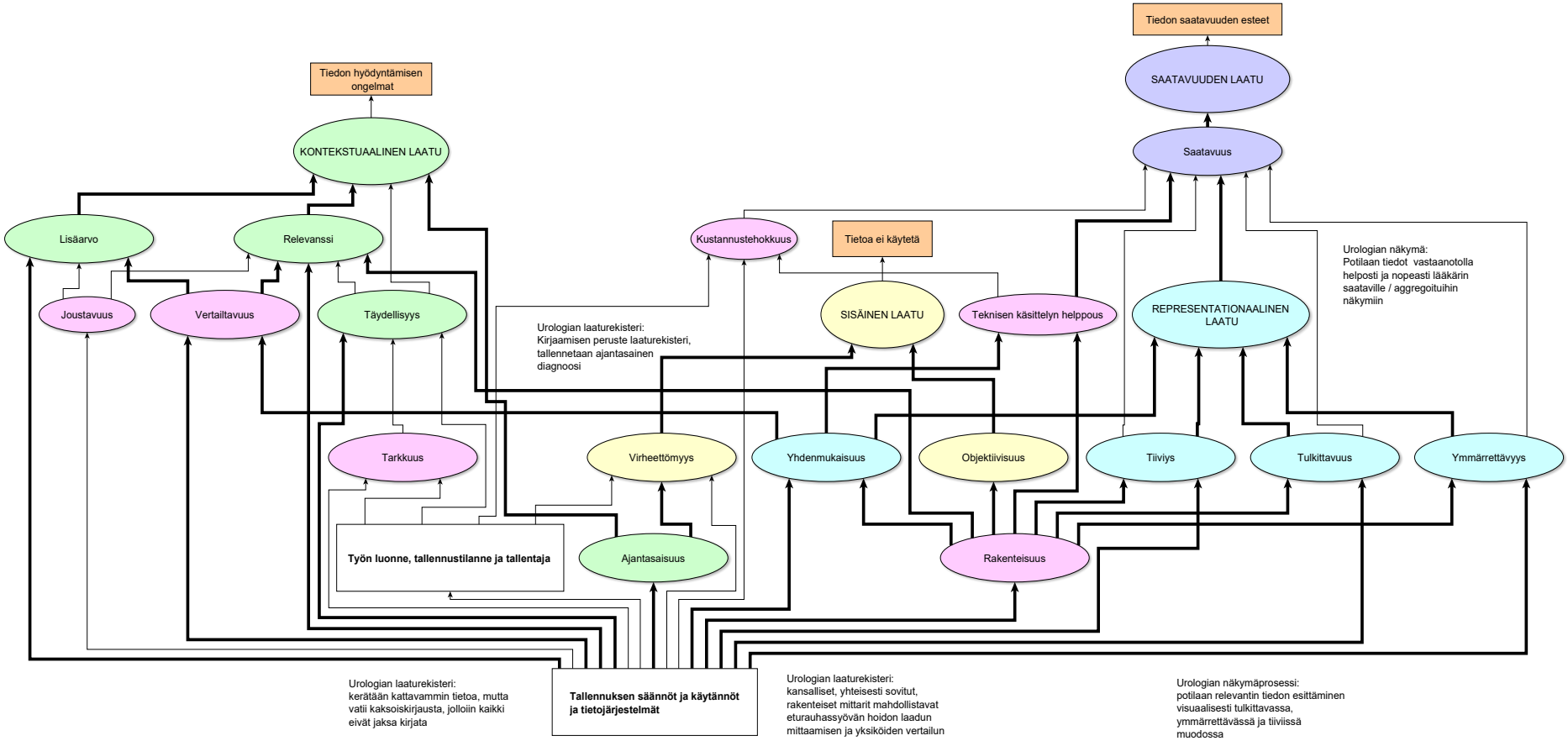
Terveydenhuollon henkilöstö tallentaa potilaan hoitoon liittyviä tietoja eri tietojärjestelmiin. Haastatteluissa puhuttiin pääasiassa lääkärin tekemästä tallennuksesta eli kirjaamisesta vastaanottotyössä. Potilastietojen tallennukseen vaikuttavat monenlaiset ja monen tasoiset säännöt, ohjeet ja tavat (kuvio 4). Useimmat haastateltavista pitivät potilastietojen kirjaamista tärkeimpänä tai perustavaa laatua olevana tiedon laatuongelmien syynä tiedon toisiokäytössä: ”Et tavallaan, vaiks tekis mitä temppuja tai ilveitä tai minkä näkösii ohjelmanpätkiä hyvänsä, niin tätä tiedollista laatuu ei pysty muuttamaan, ellei se kirjaamisen tapa muutu”. (L5)

KUVIO 4 Tiedon laatuongelmien syiden luokittelu

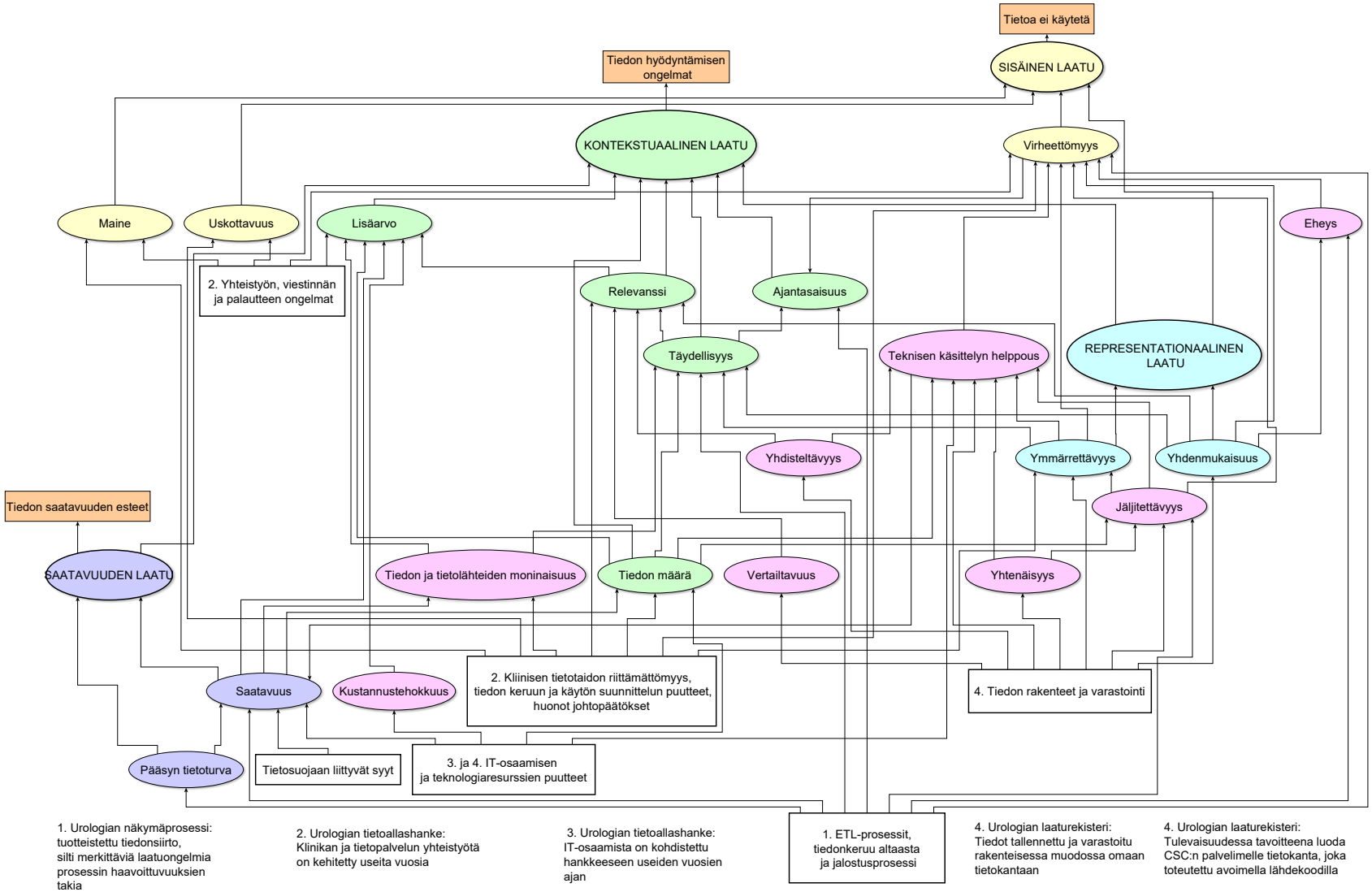
Seuraavassa käydään läpi, miten erilaiset tallennukseen vaikuttavat ***säännöt ja käytännöt*** vaikuttavat tiedon eri laatuominaisuuksiin. Näihin on luokiteltu mittarit ja sopimukset, kirjaamisperusteet, kirjaamisen ohjeistus ja käytäntö.

Mittarit ja sopimukset ovat päätöksiä siitä, mikä on merkittävää tietoa, ja miten tietoa tallennetaan, esimerkiksi mitataanko potilaan komplikaatioita ja tehdäänkö se 30 vai 90 päivää leikkauksen jälkeen. Monet haastateltavista mainitsivat yhteisesti sovitut mittarit erittäin tärkeinä tiedon laatuun vaikuttavina tekijöinä tiedon toisiokäytössä. Tiedon laatuongelmia seuraa, jos sopimukset puuttuvat. Tieto ei silloin ole yhdenmukaista **(puuttuvat sopimukset → yhdenmukaisuus)**. Tiedon yhdenmukaisuuden kannalta mittareita ja sopimuksia voidaan pitää välttämättöminä, ja juuri mittaamisesta sopiminen on ollut ensisijaista kansallisessa eturauhassyövän laaturekisterihankkeessa, jolla on tiivis yhteys urologian klinikan tietoallashankkeeseen.

Tällä hetkellä hoidon laadun ja vaikuttavuuden seurannassa pystytään yleisesti seuraamaan pääasiassa kovia mittareita, kuten kuoliko potilas, palasiko hän takaisin sairaalaan tai tehtiinkö uusintaleikkaus. Kustannusvaikuttavuusmittauksessa palveluntuottajia mitataan suoritemäärien, talouslukujen, haittatapahtumien ja osastolle



KUVIO 5 Tiedon kirjaamisvaiheessa syntyvät toissijaisen käytön laatuongelmat ja niiden syyt.
Selitteet: Ellipsi kuvaa laatuominaisuutta tai -ulottuvuutta, jota laatuongelmat koskevat. Nuoli kuvaa laatua heikentävää vaikutusta ja sen suuntaa. Paksummalla viivalla merkitty nuoli merkitsee urologian hankkeessa kehitettyjä ratkaisuja, jotka parantavat tiedon laatuominaisuutta tai -ulottuvuutta. Värien selitteet ks. kuvio 3 (s. 37).



KUVIO 6 Tiedon varastoinnin, jalostamisen ja käytön yhteydessä syntyvät toissijaisen käytön laatuongelmat ja niiden syyt. Selitteet: Ellipsi kuvaa laatuominaisuutta tai -ulottuvuutta, jota laatuongelmat koskevat. Nuoli kuvaa laatua heikentävää vaikutusta ja sen suuntaa. Numeroin merkityt tekstikentät selittävät, miten urologian hankkeessa kehitetyin ratkaisuin pyritään parantamaan tiedon laatua. Värien selitteet ks. kuvio 3 (s. 37).

palaamisen raportoinnilla, mikä ei suoraan kerro hoidon vaikuttavuudesta. Esimerkiksi osastolle palaaminen ei sovellu eri yksiköiden vertailuun, sillä se on riippuvainen yksiköiden toimintatavoista. (Sitra, 2019.) **(mittarit → yhdenmukaisuus → vertailtavuus.)** Pehmeämmät mittarit, joilla voisi mitata potilaan subjektiivista kokemusta ja jotka ovat terveystaloustieteessä tärkeä osa hoidon vaikuttavuuden arviointia, ovat käytössä vain harvalla klinikalla **(mittarit → täydellisyys → relevanssi)**. Elämänlaatumittareista ei ole olemassa laajempaa sopimusta, vaan ”siel on nyt liikkeellä niissä vähissä projekteissa mitä on niin sekalaisii mittareita, ja sit ei oo sovittu sitä, et millä aikaskaalalla niit mitataa” (L5) **(mittarit → yhdenmukaisuus)** (ks. myös Ville Äärimaa Sitran (2019) raportissa).

Koska potilastietojärjestelmiin tallennettavan tiedon ei koeta riittävän hoidon laadun mittaamiseen ja tiedolla johtamiseen, klinikat keräävät erikseen tietoa näihin tarkoituksiin. Urologian klinikalla oli aiemmin kokeilussa kaupallinen, BCB Medicalin toteuttama laaturekisteri, mutta sitä ei koettu omaan toimintaan sopivaksi joustamattomuuden ja siinä käytettyjen mittareiden vuoksi. Tuotteesta päätettiin luopua ja omaa sovellusta alettiin kehittää tietoallashankkeessa. Hankkeessa käytettävät hoidon laadun mittarit on kehitetty Suomen Urologiyhdistyksen jäsenistön yhteistyönä. Samat mittarit ovat käytössä THL:n kansallisessa eturauhassyövän laaturekisterihankkeessa.

Sopimusten puute ei ole enää eturauhassyövän hoidon laadun mittareiden osalta TYKS:in urologien mukaan merkittävä ongelma. Haastateltu urologian erikoislääkäri korostaa yhdessä sovittujen mittareiden laajaa merkitystä urologian tietoallashankkeessa: mittarit ja sopimukset mahdollistavat eri yksiköiden ja erilaisten hoitojen arvottamisen, sen mikä on kustannustehokasta hoitoa **(mittarit → lisäarvo)**. Tämä on erityisen tärkeää nykyisessä kansallisessa tilanteessa, kun terveydenhuollon kustannukset nousevat, soteuudistus on tulossa, ja hoitoa pyritään keskittämään. Eräs haastateltava kuitenkin kritisoi eturauhassyövän laaturekisteriä siitä, että siihen on valittu liikaa käsin kirjattavia asioita: ”Jotta se niinku arjessa toimii, niin se pitää olla keskeiset asiat, mutta ei ihan kaikkee. Jostain täytyy karsii.” Haastateltava viittaa siihen, että vain osalla lääkäreistä on motivaatiota kirjata vielä lisää asioita, jolloin kerätyn tiedon laatu on erittäin huono **(mittarit → työmäärä → tallentajan motivaatio → täydellisyys + kustannustehokkuus[3]).**

Jos tallennettaville tiedoille on olemassa yhtenäinen tietomalli, mitä tallennetaan ja missä muodossa, tieto saadaan helpommin ja nopeammin käyttöön, koska sen yhdenmukaistamiseen ei tarvitse käyttää niin paljon aikaa ja vaivaa. Oman sairaalan sisällä yksittäisissä tutkimusprojekteissa ”sottaisen” tiedon puhdistaminen on mahdollista, mutta suurempien aineistojen kohdalla se olisi niin hidasta ja vaivalloista, että se ei olisi enää kustannustehokasta **(mittarit → yhdenmukaisuus → teknisen käsittelyn helppous → kustannustehokkuus → saatavuus).**

Koodistot ja luokitukset ovat lääketieteessä keskeisiä sopimuksia. Laajasti käytettyjä koodistoja ja luokituksia on esimerkiksi laboratorioarvojen, diagnoosien ja toimenpiteiden kirjaamiseen. Tietoja on mahdollista kirjata eri tarkkuustasolla: ”nehän on nelinumeroisii niinku isoja numerosarjoja, joissa on niinku vaan viimesen numeron ero ja se tarkottaa solukuvassa tai jossaki muutosta tai erilaisuutta” (L1), mutta silti esimerkiksi syöpätautien osalta diagnoosi on ”ohut tieto siitä, mitä potilas sairastaa” (L5). Se ei kerro taudin vaikeusastetta tai vaihetta **(koodistot + kirjaaminen → tarkkuus → täydellisyys).**

[3] Ajatuksena on, että jos kirjataan paljon yksityiskohtaista tietoa, se vie paljon lääkärin työaikaa, mutta kerääminen ei ole kustannustehokasta, koska kerätty tieto on huonolaatuista.

Toiminnan luonne vaikuttaa *kirjaamisperusteisiin*. Potilastietojärjestelmien osalta kirjaamisperusteita ovat erityisesti potilaan hoito, oikeusturva ja talous. Erikoissairaanhoidossa tiedon puuttuminen ei ole merkityksetöntä, ”meillähän tänne kertyy niinkun dataa sitä mukaa, kun potilaat tulee kipeeks, tälleen niinku karkeesti sanottuna. Et keskimäärin kukaan näist potilaista ei oo silleen, et he käy tääl vaik viikon välein jossain tietys testissä” (I2). Tiedon puuttuminen voi vastaavasti johtua siitä, että henkilö ei ole sairastunut, tai että loppuvaiheessa olevaa sairautta ei enää kannata hoitaa **(kirjaamisperusteet → täydellisyys).** Ongelmat toisiokäytössä juontuvatkin usein siitä, että tiedot on kerätty hoidon näkökulmasta, jolloin tarkoituksena on, että lääkäri lukee tai kommentoi yhden potilaan tietoja kerrallaan. Tieto soveltuu tällöin usein huonosti retrospektiiviseen analyysiin **(kirjaamisperusteet → relevanssi).** Potilaan terveystietoja käytetään myös talouspuolella, mistä ongelmallisimpana eräs haastateltavista lääkäreistä pitää diagnoosin merkitsemistä, jotta asuinkuntaa voidaan laskuttaa. Käynnin laskutusperuste voi olla esimerkiksi eturauhassyöpä. Se näkyy potilasrekisterissä diagnoosina, vaikka potilas olisi jo parantunut **(kirjaamisperusteet → ajantasaisuus).** Tällöin diagnoositietoon ei voi luottaa **(kirjaamisperusteet → ajantasaisuus → virheettömyys).**

Urologian hankkeessa visuaalisen aikajananäkymän käyttäjät ovat riippuvaisia potilastietojärjestelmien tiedosta, joita koskevat edellä mainitut laatuongelmat. Laaturekisteritallennus sen sijaan on suunniteltu juuri tiedolla johtamisen ja tutkimuksen tarpeisiin, joten kirjaamisperusteet eivät vaikuta tiedon laatuominaisuuksiin heikentävästi, kun tietoa käytetään näihin tarkoituksiin.

Eri klinikoilla kirjataan ”erilaisia asioita eri logiikalla”, ja kirjaaminen myös muuttuu ajan myötä. Tieto ei ole ”homogeeninen massa”, vaan ”se riippuu niin paljon niist klinikoista, ja kirjauskäytännötki muuttuu ajan myötä, nii se on tavallaan vähän niinku elävä organismi se data sinäns**”** (I3) **(kirjaamiskäytännöt → yhdenmukaisuus.)** Haastateltu tietoarkkitehti painottaa, että *kirjaamisen ohjeistamisella* voidaan huomioida tiedon käyttö hoidon laadun arvioimiseen jälkikäteen eri näkökulmista **(kirjaamisohjeet → relevanssi, kirjaamisohjeet → joustavuus)**. Ohjeistuksen suunnittelussa tulisi hänen mukaansa käydä vuoropuhelua klinikoiden ja kirjaajien kanssa, koska joissain tilanteissa yleinen ohje ei välttämättä toimi. TYKS:issä tietohallintoylilääkäri vastaa kirjaamisen ohjeistuksesta.

Kirjaamisen käytäntö on analyysissä oma alakohtansa, koska totutut tavat ja vakiintuneet käytännöt voivat vaikuttaa kirjaamiseen yhtä lailla kuin ohjeistus. Kirjaamisen tavat voivat noudattaa tai olla noudattamatta virallista ohjeistusta **(kirjaamiskäytännöt → yhdenmukaisuus),** joten kun tietoarkkitehti ”luo tiedolle ymmärrystä” (I3), hänen on kysyttävä kirjaamisen tavasta suoraan niiltä, jotka kirjaavat **(kirjaamiskäytännöt → ymmärrettävyys).** Kaikkea hoitopäätösten taustalla olevaa tietoa ei ole tapana kirjata **(kirjaamiskäytännöt → täydellisyys)**. Urologisten syöpien osalta moni haastateltava mainitsi TNM-luokan[4] tietona, jota ei tallenneta tietojärjestelmiin ”vaan se syntyy […] urologin tai ammattilaisen sormen ja korvain välissä”(L6).

> tieto on kliinikon mielessä mutta hän ei sano sitä. Se näkyy rivien välistä sitten, kun todetaan, et hoitopäätös on tämä ja leikkaus ei todennäköisesti paranna tai jotain, niin siit voidaan päätellä, että no hänen mielestään ehkä T-luokka oli jotain tuon kaltaist, mut et sitä ei oo niinku missään eksplisiittisesti. (I3)

[4] Lyhenne TNM tulee sanoista tumor (kasvain), node (imusolmuke) ja metastasis (etäpesäke), ja kertoo kasvaimen koosta ja syövän leviämisestä läheisiin imusolmukkeisiin tai kauemmas elimistöön (UICC, 2020).

Monet haastateltavista mainitsivat ***potilastietojärjestelmät*** keskeisenä tiedon laatuongelmien syynä. Potilastietojärjestelmillä viitataan tässä kaikkiin niihin klinikoilla käytössä oleviin tietojärjestelmiin, joihin tallennetaan potilastietoa. Se, mitä tietoa ja *missä muodossa potilastietojärjestelmät sallivat tallentaa*, vaikuttaa suuresti tiedon laatuun. Suuri osa potilastietojärjestelmiin tallennetusta tiedosta on rakenteetonta tekstiä, joka on ”saneluissa […] monin eri tavoin ilmastu” (I3) **(sallittu tallennuksen muoto → rakenteisuus → yhdenmukaisuus)**. Lähes kaikki haastateltavat mainitsivat tiedon rakenteettomuuden tiedon laatuongelmien syynä. Haastatellun urologian ylilääkärin ja tietoarkkitehdin mukaan suurin ongelma tiedon käytössä lääkärien näkökulmasta juontuu juuri tästä rakenteettomasta, vapaamuotoisesta tiedosta, jota on pakko käyttää rakenteisen tallentamisen puutteiden vuoksi:

> se ei oo niinkun hallittavissa, siihen tarttis jotain dataminingia käyttää, ja se ei oo realistista käyttää, ja se on usein huonosti määriteltyä eli tämmöstä subjektiivista ja tällasta, ja sen takii sen datan täytyy olla strukturoidusti kerättyä, jollon se on yksselitteistä. Usein se on käyttökelvotonta dataa, jos ei se oo struktuoitua. (L8) **(rakenteisuus → objektiivisuus + ymmärrettävyys + relevanssi + (teknisen käsittelyn helppous → saatavuus))**

Joissain tapauksissa saman tiedon (esim. tupakoiko potilas) voi tallentaa sekä rakenteisesti että rakenteettomassa muodossa, mutta usein potilastietojärjestelmän muoto ei salli sitä:

> on pseudorakenteisii tietoja, et on niinku se tieto ittessään on rakenteista mut ei ole vaan mitään tämmöstä Excelin kaltasta systeemii, mihin sitä kirjattais, et yks on vaikka verenpaine. Sairaalas vois äkkiä kuvitella, et on joku semmonen taulukko ja jotain on, mut se suurin osa verenpainetiedoista on kuitenki vaan saneltuna tekstinä. (I2) **(sallittu tallennuksen muoto → rakenteisuus)**

Saneluissa samat asiat on ilmaistu monin eri tavoin ja narratiivisen tekstin määrä on valtava **(rakenteisuus → yhdenmukaisuus → tiiviys)**. Toisiokäyttöä varten tietoa on lähes aina louhittava narratiivisesta tekstistä, ja louhimiseen käytettyjen koneoppimisratkaisujen kehittäminen voi viedä aikaa riippuen käytössä olevista teknologioista ja menetelmistä **(rakenteisuus → teknisen käsittelyn helppous → saatavuus)**. Rakenteettoman tiedon laatuongelmat on urologian laaturekisterilomakkeilla ratkaistu tallentamalla tiedot rakenteisessa muodossa, jolloin rakenteettoman tiedon laatuongelmat eivät koske sitä. Aikajananäkymään käytetään kuitenkin myös tietoaltaan potilasrekisteritietoa, joten sen muoto vaikuttaa näkymän kehittämistyössä ja käytössä.

Haastateltu tietoarkkitehti pitää kiireen ohella suurimpana syynä kirjaamisen ongelmiin *potilastietojärjestelmien muotoa, niiden vanhanaikaisuutta ja sitä, että ne eivät ole kovin helppokäyttöisiä ja ketteriä*. Joskus harvoin niissä on myös *virheitä*: ”Et kyllähän potilastietojärjestelmässäki on bugeja vaik niis on isot koneistot, jotka niitä aina testaa. Et kyllähän niit virheit tapahtuu”. (I3) **(tietojärjestelmät → virheettömyys)**. Sen lisäksi, että yksittäisten järjestelmien käytettävyys ei ole ihanteellinen, potilaiden tietoja haetaan ja kirjataan moniin eri järjestelmiin, jotka ovat eri toimittajien tuotteita eivätkä keskustele keskenään. Joitakin tietoja voi tästä syystä joutua kirjaamaan moneen paikkaan **(tietojärjestelmät → kustannustehokkuus → saatavuus)**. Yhdessä kiireen kanssa sirpaleinen järjestelmäkokonaisuus hankaloittaa niiden sisältämän tiedon käyttöä potilaan hoidossa:

> jos sullon nyt kakskyt minuuttii aikaa ja potilas on sairastanu kolme tai viis vuotta niin ei niihin vanhoihin asioihin palata, koska ei niit. Sun täytys skrollata ihan hirvee määrä, ennenku sä löytäsit edes sen käynnin, mistä haluaisit puhua. (L1) **(tietojärjestelmät → tiiviys → saatavuus.)**

Vaikka potilasrekisteritiedon esittämistapa vaikuttaakin ensi sijassa potilaan hoitoon, myös toissijaisessa käytössä voi olla tarpeen tarkastella potilastietojärjestelmien tietoa. Urologian näkymäprosessissa on pyritty erityisesti parantamaan tiedon esittämistapaa tuomalla lääkärin saataville relevantti informaatio potilaan tilanteesta. Tiedon hahmottamista on parannettu visuaalisin keinoin:

> siin on, puhutaan kolmesta kaistasta, siin on kolmessa kerrokses niitä ikoneja, niin sinne ylimmälle riville hyppää semmosii urologien keskeisenä pidettyjä, joita sit sielt on esimerkiks labroista nousee tollanen, eturauhassyöpään liittyvä markkeri, PSA, mikä nostetaan sinne jo. Ja tietysti kaikki muut labrat ja muut, mitkä ei osu niihin, niin ne on sit siel alemmilla kaistoilla, et sieltä näkee niinku kaikki, mut ne olennaisimmat on nostettu vähän ylemmäs. (I4) **(tietojärjestelmät → ymmärrettävyys → saatavuus, tietojärjestelmät → tiiviys → saatavuus, tietojärjestelmät → tulkittavuus → saatavuus)**

Urologian näkymässä yleiskatsauksen potilaan tilanteeseen pystyy tekemään viidestä kymmeneen minuutissa, mikä on huima ero potilaan hoitoon käytettäviin järjestelmiin verrattuna. Kun kliinikko saa helposti ja nopeasti tarvitsemansa informaation tiiviissä muodossa esille, hän voi huomata helpommin seikkoja, jotka potilastietojärjestelmien käytön kankeuden vuoksi jäävät huomaamatta. Potilastietojärjestelmien ominaisuudet vaikuttavat tietojen kirjaamisen ohella toissijaiseen käyttöön myös silloin, kun tietoja poimitaan toissijaiseen käyttöön käsin suoraan potilastietojärjestelmistä.

Potilastietojärjestelmät eivät haastattelujen perusteella tarjoa tarpeeksi *jalostettua informaatiota* käyttäjän tueksi, kuten palautetta potilaan hoidosta tai lääkärin omasta työstä. Järjestelmistä puuttuu tietoja, joita tarvittaisiin johtajan päätöksentekoon. Muutaman haastatellun mukaan suuri syy potilastietojärjestelmiin tallennetun tiedon huonolle laadulle on se, että lääkäri ei saa palautetta työstään. Tiedon tallentajalle syntyy käsitys, että ”kukaan ei koskaan näit käytä”(L1). Palaute auttaisi kirjaajaa kehittymään työssään ja sitä kautta motivoisi häntä näkemään vaivaa:

> Sit ku ne näkee sen raportin sielt toisest päästä, et aa ku mä tein näin klik klik klik, nii mä nään suoraan täält, et kuin tyytyväisi mun potilaat on, tai kui monelle niist on tullu komplikaatio tai onks mun komplikaatio-rate vaik noussu tai laskenu, ku mä leikkaan potilait et kehitynks mä mun työssä. Ni se on se, mitä ei oo osattu tarjota viel ihan hirveesti tähän mennessä, käsittääkseni terveydenhuollossa. (I3) **(tietojärjestelmät → lisäarvo)**

Toimialuejohtajana toimiva haastateltu puolestaan kaipaa ajantasaista informaatiota johtamiseen tarvittavista ”perusperusperusasioista”, kuten palkkabudjetin kulumisesta ja toimialueen täyttöasteesta: ”jos meil on hotellitoiminto, ni hotellin jotajan keskeinen tieto on aamul se, mikä on meiän täyttöaste. Mä en saa mun toimialueen täyttöastetta mun tietokoneruudulle” (L5) **(tietojärjestelmät → täydellisyys → relevanssi.)**

Palautteen tarjoamisen hyödyistä yleensä on saatu viitteitä tietopalvelun projekteissa. Sädehoitoklinikalla hoidon intentin (kuratiiviset ja palliatiiviset hoidot) kirjaaminen on auttanut klinikkaa aiempaa paremmin arvioimaan hoitojensa tuloksia. Urologian tietoallashankkeessakin on jo alustavasti voitu vertailla leikkausten komplikaatioita kirurgikohtaisesti **(mittarit → vertailtavuus → lisäarvo)**.

Työn luonteeseen ja tallennustilanteeseen liittyvät tiedon laatuongelmien syyt eriteltiin kiireen, liian kirjaamisen ja kirjaamisen vaatiman lisätyön alaluokkiin. Muutama haastateltava puhui lääkärin *työn kiireisyydestä* syynä tiedon laatuongelmiin. Kaikkea ei yksinkertaisesti ehdi kirjata **(työn luonne ja tallennustilanne → täydellisyys)**.

> Mää luulen, et yks syy on tietysti kiire, ku on paljon potilaita ja ne pitäs hoitaa, eikä kukaan haluu keskittyy kirjaamiseen, kun haluais mieluummin keskittyy siihen potilaaseen. Et on siis resursseist pulaa, ihan kiire, paljo potilaita, (I3)

Kiire lisää riskiä tehdä kirjaamisessa virheitä **(työn luonne ja tallennustilanne → virheettömyys)**. Kiirettä voi osaltaan aiheuttaa hatara osaaminen tietokoneen peruskäytöstä, jolloin aikaa menee enemmän ja se kertautuu: ”semmonen joku ihan helppo asia vieki viis minuuttii ja jotain, niin se kertautuu äkkiä, ku on tuhansii lääkäreitä ja tuhansii käyttökertoja” (it-asiantuntija) **(tallentajat → kustannustehokkuus)**.

Jos kirjaaminen *vaatii lisätyötä*, esimerkiksi samojen asioiden kirjaamista uudelleen eri tietojärjestelmiin (ns. kaksoiskirjaaminen), kerätyn tiedon epätäydellisyys on erityinen riski. Erään haastatellun lääkärin arvio on, että jos on pakko kaksoiskirjata, vain noin puolet lääkäreistä on niin tunnollisia, että kirjaavat, jolloin kerätty tieto on käytännössä arvotonta **(lisätyö → tallentajan motivaatio → täydellisyys → relevanssi)**. Ylimääräinen vaiva on ongelma myös urologian lomakkeiden suhteen, koska kliinikot eivät jaksa tehdä kaksoiskirjaamisen vaatimaa lisätyötä:

> No, ongelmat varmaan liittyy siihen, et et se tiedon peittävyys tai kattavuus on niinkun huono. Ja se liittyy ihan siihen, että se on ihan tota toi, vaatii sen kaksoiskirjaamisen, jolloin kliinikolle se on lisätyötä, ja ne ei jaksa sitä tehdä. (L6) **(lisätyö → tallentajan motivaatio → täydellisyys.)**

Yksi haastateltava korosti, että kirjaaminen vie jo tällä hetkellä niin paljon työpanosta, että ”pitäis löytää pikapikaa semmosii tapoja kirjata, jotka käy nopeammin mutta kertoo täsmällisemmin niistä asioista, jotka ovat oleellisia” (L5).

Tallentajan sisäiset tekijät, joilla oli vaikutusta tiedon laatuun, luokiteltiin seuraaviin luokkiin: motivaatio ja asenne, tietämättömyys, kieliongelmat ja tietotekniset taidot. Pessimistinen *asenne* kirjaamista kohtaan on yleistä lääkärien keskuudessa ja on syynä tallennetun tiedon laatuongelmiin. Kun he eivät ole nähneet kirjaamisen hyötyjä,

> ni se on tehny niistä tallentajista ja sen datan tuottajista sen verran pessimistisiä, että jotkut saattaa valita vähän tyyliin niinku diagnoosilistasta päällimmäisen vaihtoehdon eikä viitti oikeesti miettii sitä (L1) **(tallentajan asenne → virheettömyys + tarkkuus)**.

Kyse voi olla myös tunnollisuuden puutteesta, koska kaikkea ei ole pakko kirjata, niin kaikki eivät niin tee **(tallentajan asenne → täydellisyys)**. Lääkärit haluavat kirjata, ”kunhan ne näkee sen motivaation ja sen syyn kirjata jotain vaiks se ois hiukan työlästä” (I3). Motivaatio voi syntyä siitä, että kliinikot ovat itse vaikuttaneet siihen, mitä tietoja seurataan, ja siitä, että he saavat palautetta omasta työstään. Tietojen käyttö omaan tutkimukseen voi myös avata silmiä näkemään, mikä tallennettujen tietojen todellinen laatu on, ja mitä hyötyjä kirjaamisesta voisi olla.

Syynä kirjatun tiedon virheisiin on ”harvemmin *tietämättömyys*, mutta sitäkin sattuu”. (L6) Myös esimerkiksi huono *suomen kielen taito tai lukihäiriö* voi johtaa virheisiin tallennuksessa **(kielitaito → virheettömyys)**. *Tietotekniset taidot* puolestaan vaikuttavat siihen,

miten tehokkaasti lääkäri pystyy työaikaansa käyttämään: ”jos ei osaa käyttää jotain systeemii, siinä aikaa menee turhaan hukkaan” **(**it-asiantuntija) **(tekniset taidot → kustannustehokkuus)**.

5.2 Potilastiedon varastointi- ja jalostusvaiheisiin liittyvät tiedon laatuongelmat ja niiden syyt

Tässä alaluvussa käydään läpi analyysin tulokset potilastiedon varastointi- ja jalostusvaiheisiin liittyvistä tiedon laatuongelmista ja niiden syistä (ks. kuvio 6), kun tallennusvaiheessa kirjattua tietoa käytetään tutkimukseen ja tiedolla johtamiseen.

Potilastietojärjestelmien toissijaisessa käytössä tiedon laadun riskejä liittyy paitsi kirjattuun tietoon myös potilastietojärjestelmien taustalla oleviin ***tiedon rakenteisiin ja tiedon varastointiin***. Niihin liittyvät tiedon laatuongelmien syyt on luokiteltu tiedon määrään, moninaisuuteen, ja tietojärjestelmien pirstaleisuuteen, viiteavainten puuttumiseen ja metatiedon puuttumiseen liittyviin syihin.

Kun puhutaan big datasta, VSSHP:n tietoaltaan sisältämä *tiedon määrä* on suhteellisen pieni. Yhdistettynä tiedon muihin haastaviin ominaisuuksiin, se kuitenkin vaikuttaa tiedon laadun kontrollointiin: ”ne datamassat on niin valtavia, et ei niit pysty käymään silleen jokaista tietoalkioo ei pysty tarkistaan, et onks se nyt oikein ja näin.” (I4) **(tiedon määrä → teknisen käsittelyn helppous → virheettömyys)**.

Toissijaisessa käytössä tiedon *moninainen muoto* on suoranaisena ja välillisenä syynä tiedon laatuongelmiin. Potilastiedon lähdejärjestelmät ovat eri toimittajien tuotteita, joissa tieto sijaitsee eri muotoisena ja erilaisissa tietorakenteissa **(lähdejärjestelmät → yhdenmukaisuus).** Skeemoissa myös tapahtuu ajoittain muutoksia, joten samankin järjestelmän tieto vaihtelee. Tiedon poimiminen useista lähteistä tekee sen hyödyntämisprosessista haavoittuvan. ”Tiedon luotettavuuteen ja paikkansapitävyyteen liittyvät vaatimukset kasvavat, mitä enemmän tietolähteitä on kysymyksessä” (I7). Mikäli tietojen semanttisesta vastaavuudesta, vertailukelpoisuudesta ja yhdisteltävyydestä ei voida varmistua, tietoa ei voida hyödyntää (I7) **(moninaiset tietolähteet → yhdenmukaisuus, vertailtavuus ja yhdisteltävyys → relevanssi)**.

Sairaalassa käytettävien monien eri potilastietojärjestelmien taustalla olevat potilastietorekisterit muodostavat *pirstaleisen kokonaisuuden*. Tietoa yhdistellään hyvin monista eri lähteistä ja tietokantatauluista:

> sen järjestelmän taustalla oleva tietojärjestelmä voi olla tosi pirstaleinen. Et se voi olla tosi monen tämmösen viittausavaimen takana, et mikä tieto liittyy mihinki, ja mikä sen arvo on. Se on vähän vaikee selittää, mut siellä voi olla niinku sellanen, et tietokannan tasolla se on niinkun, siellä voi olla satoja tai tuhansia tauluja, mihin se on koostettu se koko järjestelmä. Joku toinen järjestelmä taas voi koostua, puhutaan niinku kymmenistä tauluista, et se on aika semmonen, et sen pystyy mielessäänkin ymmärtään vielä. Sit jos se koko rakennelma on semmonen iso taulujen kimara, niin siit on vaikee, ei saada semmost kuvaa, mihin se nyt on tallentunu se tieto ja miten sen saa sieltä, just sen yksittäisen lukuarvon tietylle potilaalle poimittua. Siinä voi olla monen näköstä ihme sääntöö siinä matkalla, miten sen saa sielt kaivettuu. (I4) **(tietojärjestelmän yhtenäisyys → jäljitettävyys → teknisen käsittelyn helppous → saatavuus.)**

Data science -kehityspäällikkö mainitsee esimerkkinä Qpatin (patologian järjestelmä), josta haetaan potilasnäyte, diagnoosi ja aikaleima -tyyppiset tiedot noin 30 palasessa. ”Että se data saadaan uutettua sielt sairaalan niinku lähdejärjestelmäst, se vaatii aika paljon puliveivaamista” (I2) **(tietojärjestelmän yhtenäisyys → tiedon yhtenäisyys → teknisen käsittelyn helppous → saatavuus)**.

Tiedot ovat hajallaan eri lähteissä ja niiden yhdistäminen voi olla teknisesti haastavaa tai kiinni siihen kykenevän henkilöresurssin vähyydestä **(tietojärjestelmän yhtenäisyys → yhdisteltävyys → teknisen käsittelyn helppous → saatavuus)** (VSSHP:n johtaja Leena Setälä Sitran (2019) raportissa). Suomessa potilasrekistereistä lähes aina voidaan identifioida henkilötunnuksella potilas, jonka tiedoista on kyse. Hankaluutena tietyn potilaan tiedon käytössä on sellaisten *viiteavainten puute*, joiden avulla yksiselitteisesti voidaan yhdistää potilaan eri lähteistä poimittavia, tietoja toisiinsa:

> täytyy tehdä kaikkia semmosia eri näkösiä niinku sääntöjä, jotka haarukoidaan kellonaikojen ja päivämäärien mukaan, et tää vois liittyä nyt tähän. Ei oo semmost suoraa linkitystä aina, et tämän potilaan tämä diagnoosi, et siihen liittyy nämä tiedot näistä järjestelmistä. (I4) **(tiedon yhtenäisyys ja viiteavaimet → ymmärrettävyys → teknisen käsittelyn helppous.)**

Potilaan tietojen yhdistämisessä voi viiteavainten puutteen vuoksi tapahtua virheitä **(viiteavaimet → yhdisteltävyys → teknisen käsittelyn helppous→ virheettömyys).** Jos lääkitys tai hoito ei selkeästi liity tiettyyn diagnoosiin, sitä ei välttämättä yhdistetä siihen, tai jos aikaleima on virheellinen, tieto jää yhdistymättä, jolloin esimerkiksi urologian näkymässä esitetyn tiedon ajantasaisuus kärsii **(viiteavaimet → ymmärrettävyys → täydellisyys ja virheettömyys → ajantasaisuus).**

Dokumentaation tai *metatiedon puutteita* ei kukaan VSSHP:n organisaatioon kuuluvista haastatelluista itse maininnut tärkeimpänä syynä tiedon laatuongelmiin tiedon toisiokäytössä. Kysyttäessä metatiedosta, data science -kehityspäällikkö kertoo, että aineiston mukana käyttäjälle annetaan yleistä tietoa tietolähteiden (esim. Oberon) sisällöstä ja tiedon alkuperästä: ”labratekstis vois olla vaikka, tämä teksti on peräisin onkologialta” (I2). Lisäksi tietopalvelusivuilla on luettelo siitä, mitä eri tietolähteet sisältävät. Tietopalvelun wikissä on ainakin sisäiseen käyttöön myös ”varo näitä sudenkuoppia -tyyppinen juttu” (I2.) Lisäksi tietokantahauista ja tietojen luovuttamisesta tallennetaan lokitietoa.

Kysyttäessä siitä, mitä ongelmia voi aiheutua tiedon jalostusvaiheessa, tietoarkkitehti mainitsee, että kun on paljon toimijoita ja vaiheita, ideaalisesti olisi käytössä niin sanottu data lineage -ratkaisu, joka kertoisi jokaisen tietoalkion historian, sen, kuka sitä on käsitellyt, ja mitä sille on tehty. Tällaisten ratkaisujen tekeminen ei kuitenkaan olisi helppoa (I4) **(metatieto → jäljitettävyys)**. Jo tiedon perusrakenteiden kuvauksesta olisi hyötyä. Ilman sitä tiedolle asetettujen vaatimusten täyttäminen on haastavaa. Palveluntarjoajan edustaja nostaa esille sen, että usein tiedon jalostusputkessa joudutaan toimimaan ilman lähdetietokannan kuvausta:

> Usein ne kuvaukset on olemassa mutta ne on tällasia toimittajien liikesalaisuuden piiriin kuuluvia. Kovinkaan usein järjestelmätoimittajat eivät ole halukkaita antamaan kuvauksia siitä datan muodosta siinä lähdejärjestelmässä, jolloin sitten tällasien palveluintegraattorina toimivien 2M-IT:n tai Auria tietopalvelun rooli kasvaa siinä, että meidän pitää antaa ymmärrys sille datalle, jollemme me ole saaneet uusimpia kuvauksia siitä sisällöstä. (I7.) **(metatieto → jäljitettävyys → ymmärrettävyys)**

Tällaiset metatiedon puutteet lisäävät tarvetta ”luoda ymmärrystä datalle” ja siten laadun kontrollointiin liittyvää työtä **(metatieto → ymmärrettävyys → teknisen käsittelyn helppous).** Monet eri siirto-, muokkaus- ja latausvaiheet aiheuttavat riskin, että dokumentaatio ei kulje eheästi **(metatieto + tiedon jalostusprosessi → jäljitettävyys)**. Kun tiedon jalostusprosessissa on monia siirto-, muokkaus- ja latausvaiheita ja eri toimijoita, kuten VSSHP:llä, tutkijalle tulee helposti ongelmia tietoaineiston sisällön ja siihen tehtyjen rajausten ymmärtämisessä. Rakenteistamisen yhteydessä tiedon käyttäjän on hyvä tietää, mistä järjestelmästä tiedot ovat peräisin, sillä eri järjestelmien tiedoissa ja niiden luotettavuudessa on eroa. Tietoarkkitehti kertoo esimerkkinä käyntityyppi-tiedon rajauksesta:

> varaustyyppiset käynnit monesti poistetaan, jos halutaan niinku kattoo ihan vaan fyysisiä käyntejä, pääkäyntejä niin sanotusti. Niin joissain tilanteissahan, jos sitä dataa käytetään eikä tavallaan tiedosteta, et siel on tehty tämmönen harmonisaatio tai rajaus, niin se voi aiheuttaa virhettä, koska se käyttäjä saattaa luulla, et ne on siel mukana ja se saattaa halutaki, et ne on siel mukana, niin ei vaa oo, koska yleensä niit ei haluta. (I3.) **(jäljitettävyys → ymmärrettävyys → täydellisyys + virheettömyys)**

Ja lääkitystietolähteiden rajaamisesta:

> Meillähän on lääkitystietoja apteekin järjestelmästä, sytostaattihoitojen suunnittelujärjestelmästä, sähkösist resepteistä ja osastolla tehtävistä lääkemääräyksistä. Monesta paikasta. Jos me tehdään joku raportti missä on mukana vaan yks näistä, ni jos se raportin käyttäjä ei hiffaa, et täs on muuten mukana yksi neljästä. (I3.) **(jäljitettävyys → ymmärrettävyys → täydellisyys + virheettömyys)**

Organisaation ***puuttuvat henkilö-, osaamis-, taloudelliset- ja teknologiaresurssit*** sekä organisaation kyky hyödyntää näitä vaikuttavat tiedon laatuun toisiokäytön yhteydessä. Analyysissä resurssit on luokiteltu kliinikon tietotaitoon, tekniseen ja teknologiaosaamiseen sekä teknologiainfrastruktuuriin ja teknisiin ratkaisuihin.

Haastattelujen perusteella tiedon toisiokäytössä tiivis yhteys tiedon tuottajiin on erittäin tärkeää, ja jos *lääkärin käytännön kokemusta* ei oteta huomioon, ”niin sitte joku menee pieleen” (I2). Sitä tarvitaan läpi koko tiedon hyödyntämisprosessin. VSSHP:n johtaja Leena Setälä korostaa (Sitra, 2019):

> On tunnettua, että sosiaali- ja terveydenhuollon raakadata ei ole kaikin osin luotettavaa ja että sen vuoksi voi syntyä virheellisiä analyysejä. Tiedon sisällön ja merkityksen arviointi edellyttää niiden panostusta, jotka osallistuvat tietojen kirjaamistarkoituksen ja –tavan määrittelyyn sekä niiden vahvaa osallisuutta, jotka tietävät, miten asiakas- ja potilastietojen kirjaaminen toteutuu arjen palvelutuotannossa, mitä kirjataan ja minne ja mitä merkityksellistä tietoa ei kenties lainkaan kirjata lähdejärjestelmiin.

Tiedon käytön suunnittelu on yksi kriittisimmistä vaiheista, jossa tarvitaan kliinistä tietotaitoa: ”kun tuntee sen substanssin ja tietää kirjaamistavat, niin sillon voi osata kysyä oikeita kysymyksii ja saada kuitenki jotain ihan järkevää ulos” (L5) **(kliinikon tietotaito + tiedon käytön suunnittelu → relevanssi)**. Kliinikoilla on tietoa tiedon rakenteista, esimerkiksi siitä, millaisia tietoalkioita vaikkapa sytostaattihoitoa koskeva aineisto sisältää, ja millä tavalla tiedot on järjestelmiin tallennettu:

> monta kertaa on niin, et näis järjestelmis on jotain paikkoja tai palasii, jotka näyttää hienolt, että tuollahan noi tiedot on, mut sit lääkäreitten kans ku juttelee, he kertoo, et no, ei muuten oo, kun ei me käytetä sitä tai kirjataanki se tonne. (I2) **(tiedon rakenteet + kliinikon tietotaito → ymmärrettävyys)**

Joskus tieto ”saattaa näyttää oudolta, mut sit se onkin ihan paikkansapitävää”, kun varmistetaan kliinikolta ”et onko tulokset yleensäki niinku tällasii” (I4) **(kliinikon tietotaito + tulosten tulkinta → uskottavuus)**. Potilastietojärjestelmistä poimitun tietoaineiston ymmärtämiseen tarvitaan kliinikon kokemusta. Ne on suunniteltu potilaan hoitoon ja sellaisiksi, että toinen lääkäri ymmärtää, mitä niihin on yksittäisestä potilaasta kirjattu. Vastaavasti kliinikko on tärkeä myös analyysivaiheessa. Jos kliinikoita ei ole mukana tekemässä johtopäätöksiä tuloksista, ne voivat olla vääriä.

Haastateltu palveluntarjoajan tuotepäällikkö korostaa kliinikon vastuuta tiedon käyttökohteen määrittelystä ja siitä, että tiedon käsittelijä ymmärtää tietotarpeet oikein. Kliinikko on myös vastuussa tietojen paikkansapitävyydestä ja niiden käyttämisestä vain määriteltyyn tarkoitukseen. (I7.) Kliinikoilla on potilashoidon ja potilastietojärjestelmien osaamista, mutta tiedon toissijainen käyttö vaatii etenkin tiedolla johtamisessa vielä uudenlaista tietotaitoa. ”Tääki osaaminen on vielä aika hapuilevaa, että miten huolehditaan siitä, että noukitaan vaan ne ihan tärkeimmät ja keskeisimmät tiedot laaturekisteriin” (L5) **(kliinikon tietotaito → relevanssi)**.

Jos kliinikon kokemusta ei oteta tiedon hyödyntämisprosessin eri vaiheissa huomioon, tiedon laatu kärsii. Kliinikoiden osallistuminen vaatii heiltä motivaatiota ja mahdollisuuksia osallistua kehittämistyöhön. Kuvantamistiedon hyödyntämistä hidastaa tällä hetkellä se, että annotoituja aineistoja tekoälyn opettamiseen on vasta vähän, ja tähän tarvitaan radiologeja **(kliinikon tietotaito + IT-ratkaisut ja -osaaminen → tiedon määrä → täydellisyys → relevanssi → lisäarvo)**. Lääkäreiden osallistuminen myös laajemmin erilaisiin valtakunnallisiin hankkeisiin, kuten toisiolaki, biopankkien lainuudistus ja genomilaki sekä sote-tiedon arkkitehtuurihanke, on tärkeää.

Teknisen osaamisen puutteet ovat yleisesti terveystiedon toisiokäytön esteenä. Urologian ylilääkäri tuo esille sen, että toisiokäyttöön tarvitaan kliinisen sekä IT:n korkeaa osaamista ja tiimejä, joissa on molempia:

> nää on monimutkasii haastavii projekteja, joihin Suomes on aika vähän osaamista mum mielest tällä hetkellä, siis semmosia tiimejä, joissa on sekä tätä klinikkapuolen osaamista tähän ja toisaalta tuota tuota varsinkin tää IT-osaaminen, haastavia juttuja kyllä. (L8.)

Haastatteluissa mainittiin tarpeellisina taitoina muun muassa tietoturvaosaaminen ja käyttöliittymän visuaalinen suunnittelu. IT-asiantuntijoiden osalta haastavaa on, että julkinen sektori joutuu kilpailemaan parhaista osaajista yksityisten yritysten kanssa, mihin harvalla organisaatiolla on taloudellisia resursseja. Auria tietopalvelujen henkilökunta koostuu eri alan koulutuksen saaneista henkilöistä, joilla on erilaista osaamista, mutta haastatellun biostatistikon mukaan sovelluskehityksen asiantuntijasta voisi vielä olla hyötyä tiedon laaduntarkkailussa.

2M-IT:n tuotepäällikkö painottaa, että potilasrekisteritiedon jalostamiseen vaaditaan monipuolista teknistä ja teknologiaosaamista:

> sen datan jalostusputken ja sen, niiden variaatioiden määrä lisää sen työn haastavuutta ja myöskin tällasta, pitää pystyä ylläpitämään erilaista osaamista erilaisten tietokantahakujen teettämiseen. On ne sitten JDBC-rajapintoja tai RESTiä tai mitä tahansa niitä, niin se asettaa vaatimuksia sille teknisen henkilökunnan osaamiselle. (I7.)

Tiedon toisiokäyttöön tarvitaan korkeaa IT-osaamista myös analytiikan puolella. VSSHP:llä tekoäly-, esimerkiksi neuroverkko-osaaminen, on nopeuttanut ja helpottanut yhdessä kehittyneiden työkalujen ja kasvatetun prosessointitehon kanssa tietoaltaan sanelutiedon (esimerkiksi tieto potilaan tupakoinnista) rakenteistamista toisiokäytön tarpeisiin **(IT-ratkaisut ja -osaaminen → teknisen käsittelyn helppous → saatavuus → lisäarvo)**. Myös lääkäreillä olisi hyvä olla ymmärrystä algoritmeista ja tekoälystä, vaikka tietokoneiden perusosaamisessakin on lääkärikunnalla vielä puutteita. Lääkäreiden puuttuva teknologiaosaaminen voi olla riski esimerkiksi toisiokäytön tiedontuotannossa ja tulosten tulkinnassa.

VSSHP:n tietoallasratkaisu vaatii *kattavaa infrastruktuuria ja teknisiä ratkaisuja*. Tietoaltaasta vastuussa olevan palveluntarjoajan on kyettävä Hadoop-levyjärjestelmän lisäksi ylläpitämään monenlaista teknologiaa lähdejärjestelmien tietojen siirtämiseksi tietoaltaaseen:

> Käytännös katsoen ne vaatimukset on sille, että pitää olla joku rajapinta, mistä sitä tietoa voidaan lukea, pitää olla tekniset tunnukset sinne tietokantaan, pitää olla se tiedonsiirtoyhteys, tietoliikenneyhteys sen asiakasrekisterin ja tietoaltaan välillä, ja sitte tietysti pitää olla ne tietoaltaan rakenteet kunnossa siinä mielessä et on rakennettu asiakaskohtainen hakemistorakenne, johon ne eri asiakasrekistereistä tulevat tiedot voidaan tallentaa. (I7.)

Sairaalan puolella on vastaavasti ylläpidettävä omaa arkkitehtuuria. Tällä hetkellä 2M-IT ei vielä tuo tietoa aivan kaikista potilastietojärjestelmistä altaaseen, joten tietopalvelutiimi vastaa osasta tietolatauksia. Arkkitehtuurin nykytila on ”ideaalikuvaa […] monimutkaisempi” mutta ajan myötä tavoitteena on, että ainoastaan 2M-IT tuo tietoja toisiokäytön jalostetumpiin näkymiin (I4).

Kooltaan suurten tiedostojen ja aineistojen, kuten kuvien, käsittelyyn tarvitaan runsaasti laskentatehoa. VSSHP onkin vuonna 2019 hankkinut tehokkaita grafiikkaprosessoreja tätä varten. Niitä käytetään myös sanelutiedon rakenteistamiseen neuroverkkoon perustuvan Lumbit-ohjelman avulla, jolla esimerkiksi potilaan tupakointistatuksen rakenteistaminen on aiempaan sääntöpohjaiseen tiedonlouhintaan verrattuna hyvin nopeaa ja helppoa: ”kun oli valmis malli ja tota oli valmis data, ja sitte sen algoritmin tai sen säännöstön keksiminen ei kestäny kun muutaman minuutin ja toimii tarkemmin, ku ne mihin meni ennen viikkoja siihen säätämiseen” (I2) **(IT-ratkaisut → teknisen käsittelyn helppous → saatavuus)**. Radiologian kuvien pohjalta on puolestaan kehitetty mallinnusta, jonka avulla voidaan tunnistaa, onko kuvassa syöpää tai millaisia geenimutaatioita minkäkinlaisissa kasvaimissa on **(edellinen syy-seurausketju → lisäarvo)**. Teknologian ja työkalujen avulla tietoaltaassa olevista ainutlaatuisista, retrospektiivisistä aineistoista voidaan jalostaa arvokasta tietoa:

> Et just vaikka joku harvinaisempi syöpä, et tarvitaan oikeesti kakskyt vuotta pitää katiskaa vedessä, et tulee niinku riittävästi semmosii tapauksii, mist vois tehä jotain. Ja siihen tää olis tosi hyvä, et ku ne kaikki näytteet on tuolla meiän kellarissa, ne vaan täytyy sitte preparoida ja laittaa siihen lasille ja värjätä ja skannata, se on niinku päivien työ ja siihen voi sit liittää tän kaiken datan, mitä niistä potilaista sit on karttunu, jotka on tuolla altaassa, niin vois kattoo vaikka niitä mu-

> taatiojuttuja tai jotain ihan että tän näköset syövät, ni potilaat eli näin kauan ja jotain, ku meillähän on hirvittävän pitkä seuranta. (I2) ((**tiedon määrä + tiedon ja tietolähteiden moninaisuus) + (IT-ratkaisut- ja osaaminen → teknisen käsittelyn helppous → saatavuus) → lisäarvo.)**

Analytiikan ratkaisujen kehittäminen on edellytys tietoallastiedon hyödyntämiselle. Urologian tietoallashankkeessa kehitettyä näkymää pidetään tärkeänä edistysaskeleena:

> Mut jos nyt pysytellään täs niinku hoitamisen puolella ja tutkimuksen puolella niin tota mä koen et se on yks tärkeimpiä saavutuksia, et meil on nyt se prototyyppi [aikajanasovellus]. Me voidaan tehdä siitä niinkun eri tautiryhmille, eri klinikoille ni näkymiä ja sitä kautta niinku saada se tiedolla johtaminen ja näyttöön perustuva lääketiede ni oikeesti toteutumaan. (L1) **(IT-ratkaisu → saatavuus → lisäarvo)**

Urologian hankkeessa on hyödynnetty avointa lähdekoodia, jotta järjestelmän implementoiminen onnistuisi mahdollisimman pienin kustannuksin. Tietokanta on tarkoitus viedä CSC:n palvelimelle, ”THL ylläpitää sitä, saatais kansallisesti tämmönen toivottavasti kustannustehokas tietokanta, mitä jokainen pystyis käyttämään. Siin ois helppo käyttöliittymä ja se on se niinku se pohja” (L6) **(IT-ratkaisut → kustannustehokkuus + saatavuus → lisäarvo)**.

Tekoälyn ja big datan hyödyntämisessä ennustavan analytiikan saaminen käyttöön, esimerkiksi komplikaatioiden kehittymisen ennustaminen, on tärkeää. Haastatellun tuotantopäällikön mukaan tiedossa on paljon potentiaalia, jota voidaan hyödyntää, kun käytössä on oikeat resurssit:

> Se itse tiedon hyödyntäminen on vielä aika pienimuotosta, että toivoisi, että sen hyödyntämispotentiaali olisi laajemmin hyödynnettävissä ja puhutaan tällasista uusista datapohjasista palveluista, joita nyt tämä urologian esimerkkikin osottaa, niin toivoisi, että olisi enemmän tällasia rohkeita avauksia. Ne voi perustua koneoppimiseen tai ohjelmistorobotiikkaan tai mihin tahansa keinoälyn hyödyntämiseen tulevaisuudessa, siinäki AI kehityksessä on käytännös katsoen kolme keskeistä tekijää, on se ensimmäinen vaatimus, että pitää olla dataa saatavilla, toinen vaatimus, että on tehokkaita laskentaympäristöjä, joissa sitä dataa voidaan hyödyntää ja kolmas vaatimus on tietysti se osaaminen, johon me vaikutamme sitten kouluttautumisella ja huolehdimme siitä, että meillä on osaavia asiantuntijoita Suomessa. (I7.) **(tieto/ big data + IT-ratkaisut ja -osaaminen → saatavuus → lisäarvo)**

Tietosuojaan liittyvät syyt on luokiteltu pääsyyn tietoihin, lupaprosessin vaativuuteen ja tietoturvaan liittyviin syihin. Potilastiedot ovat korkeimman suojaluokituksen tietoja. Niitä suojataan vahvasti lainsäädännöllä, ja käsittelijältä vaaditaan käyttölupa. Vaikka laki tiedon toissijaisesta käytöstä sallii rekisterinpitäjien aiempaa helpommin hyödyntää tietoa omaan käyttöönsä, 2M-IT:n tuotepäällikkö mainitsee *lupaprosessin* edelleen merkittävänä terveystiedon hyötykäyttöön liittyvänä ”vaativana ponnistuksena” (I7). Sekä lupien hakijoiden että rekisterinpitäjän on varattava lupaprosessiin aikaa ja resursseja. TYKS:in sisällä tutkimusluvan saaminen vie noin kahdesta neljään viikkoa hakemuksesta. Sitä haetaan Turku Clinical Research Center’istä tutkimussuunnitelman perusteella. Suunnitelma voidaan hyväksyä tai hylätä eettisyyden ja tieteellisen merkittävyyden perusteella **(tietosuoja → saatavuus)**.

Lupahakemuksessa nimetään tutkijat ja muut tiedon käsittelijät, joille haetaan tietojen käyttölupaa. Luvan myöntää johtajaylilääkäri. Auria tietopalvelu luovuttaa oikeudet vain niihin tietoihin, joihin on myönnetty käyttöoikeus lupaprosessissa **(pääsyn tietoturva**

→ **saatavuus)**. Tutkimuksesta poiketen organisaation oman tiedon käyttö tiedolla johtamiseen ei vaadi lupaprosessin läpikäymistä, koska se kuuluu johtajan toimenkuvaan.

Terveystietojen luottamuksellisuus ja kansalliset lainsäädännöt asettavat haasteita lääketieteen kehitykselle. Todella hyvin toimivien neuroverkkojen opettamiseen vaaditaan suuria tietomääriä, kuten annotoituja kuva-aineistoja, usein vähintään ”Suomen tai Euroopan laajuinen pläjäys”, ja ”sitä isompi pläjäys, mitä harvinaisempi tauti”, *eikä niitä ole vapaasti saatavilla* vahvan yksityisyyden suojan vuoksi (I2) **(tietosuoja → saatavuus → tiedon määrä → lisäarvo)**.

Terveydenhuollon tiedot ovat usein eri rekisterinpitäjien hallinnassa ja sijaitsevat paitsi erilaisissa tietokannoissa myös fyysisesti ja teknisesti erillään. VSSHP:n tietoallas on mahdollistanut Läntisen syöpäkeskuksen eli kolmen keskussairaalan tietojen yhteismitallistamisen ja tuomisen yhteen, ja erilaisten syövän hoidon laatutietojen julkaisemisen. Suuri puute vielä on, että perusterveydenhuollon tietoja ei ole vasta kuin kahdesta kunnasta (Salo ja Paimio-Sauvo). Esimerkiksi urologian tietoallashankkeessa potilaan hoitopolusta puuttuu lähes kokonaan erikoissairaanhoidon ulkopuoliset tiedot **(pääsy tietoihin → saatavuus → tiedon ja tietolähteiden moninaisuus → täydellisyys)**.

TYKS:issä tiedon toisiokäyttöä ei haastattelujen perusteella ole koettu erityisesti *tietoturvaongelmaksi*, mikä todennäköisesti liittyy siihen, että tietoallas on toteutettu sillä periaatteella, että ”se on potilastietojärjestelmän osa ja sen kanssa saman tietosuojaselosteen alainen” (L1) ja haastattelut koskivat organisaation sisäistä tiedon käyttöä. Tietoa ei luovuteta sairaalan ulkopuolelle, vaan tutkijat tuodaan sairaalaan. Toisiokäytön harrastaja ei myöskään pääse käsiksi suoraan tietoaltaan tietoon, vaan hänelle tehdään erillinen aineisto:

> heille erotetaan sitte näil tutkijoil tai tiedolla johtajille ni oma kohortti heidän tarpeittensa mukaan, mis on ne potilaat sitte tai potilaan tiedot, mitä siinä hankkees tarvitaan. Ja se otetaan tietovarastoon erilliseksi ja korjataan, paikataan, mitä se sit tarvii ja annetaan sit siihen tietoturvallisel-le alustalle käsiteltäväksi, tai tehdään jopa ajot sitten, jos asiakas haluaa. Ja tää oli se meidän perusprinsiippi ja sitä lähdettiin sit tekemään. (L1) **(pääsyn tietoturva → saatavuus)**

Tiedon jalostamisen vaiheista johtuviksi syiksi luokiteltiin tiedon käytön suunnittelun puutteet, tiedonkeruun käytännöt, tiedon siirto-, muokkaus- ja latausvaiheet sekä analyysin ja johtopäätösten laatu. *Tiedonjalostusprosessin suunnittelu* ja vaatimusmäärittely sisältää ihanteellisessa tapauksessa tiedon tarkan käyttökohteen kuvaamisen, minkä pohjalta jalostusprosessin vaiheet suunnitellaan. Kliinikon näkökulmasta tärkeä vaihe on käyttökohteen määrittelyn lisäksi *tiedonkeruu*. Kysyttäessä tiedon laatuun liittyvistä epävarmuustekijöistä urologian ylilääkäri painottaa tiedonkeruun huolellista suunnittelua tutkimuksessa ja tiedolla johtamisessa: ”jos se olis suunniteltu huonosti, niin siinä vois olla paljonki kyllä ongelmia”(L8). Haastattelussa ei puhuta tarkemmin näiden ongelmien laadusta, mutta toisen haastattelun perusteella kliinikon osaamista tarvitaan relevanttien kysymysten esittämiseen suunnitteluvaiheessa, mikä on puolestaan edellytyksenä tulosten järkevyydelle **(kliininen osaaminen → suunnittelun laatu → relevanssi → lisäarvo)**. Esimerkiksi oikeiden tietolähteiden valinta (laaturekisteri vai tietoallas) on olennaista, jotta ei tehdä virheellisiä johtopäätöksiä (Sitra, 2019) **(toisiokäytön suunnittelun ja analyysin laatu → johtopäätösten virheettömyys)**. Tietoarkkitehdin mukaan tiedolla johtamisessa voitaisiin hyötyä siitä, että käytettäisiin useampia tietolähteitä: ”ihan kaikkii tietoo ei oo viel tän kaiken toisiokäy-

tön niinku isossa kuvassa osattu tuoda yhteen”(I3) **(toisiokäytön suunnittelun laatu → tiedon ja tietolähteiden moninaisuus → lisäarvo)**.

Potilaiden tietosuojan on koettu VSSHP:ssä parantuneen jalostetumman tutkimusaineiston tarjoamisen myötä. Aiemmin palveluntuottajalta (2M-IT) pyydettiin henkilötunnuksia potilaista, joilla oli vaikkapa eturauhassyöpädiagnoosi, ja tunnusten perusteella tutkija itse haki tietoja suoraan potilastietojärjestelmien hoitonäkymän kautta, missä on näkyvissä laajasti henkilö- ja terveystietoja. Tällainen tiedonkeruu on vähentynyt huomattavasti uudessa prosessissa. VSSHP:ssä tarjotaan usein henkilötunnuslistan mukana potilaasta valmiina muitakin tietoja (esim. diagnoosit ja toimenpiteet) ja johdettuja suureita, jolloin ”ihan kaikkee hänen terveystietoaan ei tarvii siinä ihmisten katsella”(I3). Tekstitietojen saaminen altaaseen helpotti ja nopeutti useiden tietojen (esimerkiksi verenpaine) poimimista, kun niihin päästiin helpommin käsiksi automaattisilla menetelmillä. Uusi prosessi on paitsi parantanut potilaan tietosuojaa myös vähentänyt tietojen poiminnassa tapahtuvia virheitä ja säästänyt aikaa potilastietojärjestelmistä tapahtuvaan manuaaliseen tiedonkeruuseen verrattuna (I2, I4) **(toisiokäytön tiedonkeruun laatu → pääsyn tietoturva + virheettömyys + saatavuus.)**

> perinteisesti homma menee niin, että joku saa idean, voi vitsi et ku ois kiva tutkii jotain tiettyy syöpää, alanpa keräämään siihen aineistoa sitä mukaa kun semmonen potilas kävelee ovesta sisään, ja sit kirjataan ylös juttuja mut nytten sitte voidaan jo ammentaa sieltä puulista, siellä kun on viidelttoista vuodelta sitä dataa tai enemmänki jossain tapauksessa, niin voidaan kaivaa kaikki semmoset harvinaiset tapaukset ja sit tutkii niitä. (I2.)

Lähtökohtaisesti tiedon käsittelyssä on aina virheen mahdollisuus. Monista eri potilastietojärjestelmistä toissijaiseen käyttöön jalostettavalla tiedolla on monia erilaisia *siirto-, muokkaus- ja latausvaiheita*, joita hoitavat eri toimijat. Tämä lisää tiedon jalostusputken haastavuutta. Haastateltu tietoarkkitehti pitää itsestään selvänä, että tieto muuttaa muotoaan, prosessi muistuttaa ”rikkinäistä puhelinta” ja vaatii paljon silmälläpitoa (I3) **(siirto-, muokkaus- ja latausvaiheet → eheys)**. Tietopalveluiden biostatistikko painottaakin tiedon jalostusprosessin riskialttiutta tiedon laadulle ja sitä, että laadun ajantasainen hallinta on vaikeaa:

> täs on useempi tämmönen vaihe, missä voi mennä jotaki, tuntuu kauheen niinku negatiiviselta puhua, että menee pieleen, että niinkun sellasii vaiheita, missä pitäis osata varautua vähän kaikkeen […] datamassat on niin valtavia, et ei niit pysty käymään silleen jokaista tietoalkioo, ei pysty tarkistaan et onks se nyt oikein ja näin, mut et tähän pitäis jotenki pystyä, pystyä sitten vaikuttaan tai ees jotenki niinku seuraan sitä, et miten se prosessi siinä sitten menee. (I4) **(tiedon rakenne ja varastointi + tiedon siirto- ja muokkaus- ja latausvaiheet → jäljitettävyys → virheettömyys)**

2M-IT:n tuotepäällikkö pitää haasteellisena lähdejärjestelmien tiedon lukemisen tapoja, päivitysrutiineja ja tiedon lataamista siten, että ensisijaista, tuotantokäyttöä ei vaaranneta. Tietoallaslatauksessa hänen mukaansa ovat haasteina tiedon metarakenteen muutokset ja latauksen toimiminen vasteaikavaatimuksen mukaan. Haastateltu biostatistikko (I4) pitää juuri tätä vaihetta erityisen riskialttiina tiedon laadun kannalta.

Lähdejärjestelmät ovat erilaisia, ja niistä tuodaan tietoa eri ajanjaksoilta. Useimmista järjestelmistä päivän tiedot ajetaan altaaseen seuraavana yönä, toisista uudet tiedot viedään vaikkapa parin kuukauden välein, ja toisaalta osa tiedoista tulee altaaseen ”tipoit-

tain", lähes reaaliajassa. Yleensä lähdejärjestelmistä voidaan tuoda muutokset inkrementaalisesti, mutta joistain järjestelmistä se ei onnistu, vaan niistä on ensin tuotava koko tietomassa "johonkin pisteeseen", jossa vasta poimitaan uusi tieto ladattavaksi altaaseen. Inkrementaalisissa latauksissa olevia puutteita on hankalia huomata **(tiedon siirto-, muokkaus- ja latausvaiheet → täydellisyys + ajantasaisuus)**. Lähdejärjestelmien ETL-prosessit ovat siis erilaisia, ja tietoon ja sen rakenteisiin tehdään myös ajoittain muutoksia:

> mä en muista, mistä järjestelmästä oli kyse, mutta siellä saatiin XML-muodossa sitä, ja sitte se oliki se XML:n skeema jotenki muuttunu, eli sitä ei ollu vaan kerrottu meille, sitten ne tietolataukset aina öisin feilas sitte sen takia, kun se oli muuttunu se skeema, et meidän päässä tietoallas ei osannu prosessoida sitä, et sielt tuli neliönmallist palikkaa, kun laitettiin pyöreeseen reikään, et siel tuli niinku väärässä muodossa se tieto. (I4.) **(tiedon siirto-, muokkaus- ja latausvaiheet + yhdenmukaisuus → ajantasaisuus + täydellisyys + eheys)**

Paitsi palveluntarjoajan osuudesta, ongelmat tietolatauksissa voivat johtua teknisistä syistä VSSHP:llä, kuten siitä, että toinen prosessi on estänyt latauksen, tai esimerkiksi tila tai muisti loppuu.

Kun käytetään paljon ja monenlaista tietoa, tietojen harmonisoinnissa ja yhdistämisessä voi tapahtua tiedon laatuun vaikuttavia virheitä: "Tai jos me yritetään harmonisoida sitä, et siin on joku ongelma, et jotku sarakkeet ei mee niinku nätisti, päivämäärät on toisessa jotenki eri formaatissa ni, niitku yhdistetään, ni kyllähän siin voi tapahtuu virheitä"(I3) **(tiedon muunnokset + yhdenmukaisuus → virheettömyys)**. Lisäksi "keinoälyhommissa" voi tapahtua virheitä, esimerkiksi rakenteisen tiedon algoritminen poimiminen sanellusta tekstistä (esim. tupakointistatus) ei tuota sataprosenttisen oikeellista tulosta **(tiedon muunnokset → virheettömyys)**.

Urologian tietoallashankkeen ongelmana on, että ensisijaiseen käyttöön potilaan hoidon lisälaitteelle vaadittavaa CE-merkintää urologian prototyypille on mahdoton saada sairaalan omin voimin, sillä potilastietojärjestelemistä saatu tieto ei kulje prosessissa potilasturvallisuuden vaatimalla tavalla eheästi. Tiedon eheyttä toissijaisessa käytössä urologian tietoallashankkeessa parantaa kuitenkin se, että prosessi on tuotteistettu, ja sitä voidaan jatkuvasti kehittää.

Moni haastateltavista korostaa, että on riski tehdä *johtopäätöksiä* ilman kliinikkoja: "ne on siihen perehtyny, et kyl niil saattaa olla jotain omaa raportointitietoo tai taustaa, mitä ne on hankkinu tai vertailutietoo muist sairaaloista" (I3). Jos tehdään huonoja johtopäätöksiä ja jokin menee pieleen, voi syntyä kielteinen asenne tietoaltaan hyödyntämistä kohtaan ja se voi saada huonon maineen:

> se voi kääntyy vähän siihen, et no, mitä järkee tämmöses tietoallastouhussa tai ihan tyhmää tehdä tollasii rekisteritutkimuksii, ku ei niist oo mitään hyötyä ja niin poispäin, tulee niinkun negatiivinen asenne niitä kohtaan […] plus sitte tietysti kahvipöydässä leviää se homma, ja muutki on sitte, et ei mekään sitte edes yritetä, muutenki on kiire ja stressi näitten ihan tavallisten töitten kanssa. (I2) **(johtopäätösten laatu + viestintä → maine)**

Toisaalta myös tieto onnistuneista tiedolla johtamisen projekteista leviää puskaradiossa, ja sitä kautta moni osaa pyytää aineistoa tietopalvelulta.

Kansallista vertailutietoa ei voida tulkita pelkästään kansallisten mittarien pohjalta. VSSHP:n johtaja Leena Setälä (Sitra, 2019) korostaa, että tietoaltaan tiedosta johdetut kansalliset kustannusvaikuttavuusmittarit eivät yksinään riitä palvelujärjestelmän toimivuu-

den arviointiin. Niiden tulkinnassa olisi lisäksi otettava huomioon potilaaseen ja palvelujärjestelmään liittyviä tarkempia tietoja, sillä palvelujärjestelmän rakenne ja potilaiden taudin vaikeusaste tai hoitotasapaino vaikuttaa lukuihin eri maakunnissa eri tavoin (Leena Setälä Sitran (2019) raportissa) **(tiedon ja tietolähteiden moninaisuus → täydellisyys → relevanssi → lisäarvo.)**

Viestintään liittyvät syyt liittyivät yhteistyöhön, viestintään ja palautteeseen. Kun mukana on monia tahoja, toimijoita ja vaiheita, *saumaton yhteistyö ja viestintä* on tärkeää, mikä tulee esille useissa haastatteluissa. VSSHP:n oma tietopalvelu mahdollistaa IT-osaajien ja kliinikoiden tiiviin yhteistyön. Tällainen läheinen yhteistyö on edellytys sille, että urologian hanke ja muut samantyyppiset hankkeet voivat onnistua **(yhteistyö ja viestintä → lisäarvo)**. Haastatellun urologin mukaan TYKS:issä tietopalveluyksikön asiantuntijoilla ja urologeilla on mennyt yhteisen kielen löytymiseen ”pari vuotta” (L6). Läheisen yhteistyön on mahdollistanut se, että ”ollaan täs saman talon sisällä ja vielä fyysisesti viereisissä rakennuksissa, et pystyy aika silleen helposti järjestään kokouksia, et mis voidaan sit miettiä niitä kehityssuuntia”(I4).

Tietoallastiedon hyödyntämisprosessin alkuvaiheessa IT-asiantuntijoiden piti tehdä aika paljon ”salapoliisityötä”, kun jokin uusi potilasrekisteri tuli tietoaltaaseen (I2). Kun tieto kulkee läpi useita vaiheita, ongelmatilanteiden selvittelyssä on tärkeää ottaa yhteyttä oikeaan henkilöön tietopalvelussa tai 2M-IT:llä. Palaute tiedon laatuongelmista, kuten tietolatauksen puuttuvista riveistä, on tärkeä osa prosessia. Kliinikoiden palaute urologian näkymästä on tietopalvelulle tärkeää tiedon laadun seurannan näkökulmasta: ”saadaan kuitenki pientä semmosta niinku feedbackiä siinä koko ajan, et okei, kyl meiän dataki on varmaan suurinpiirtein aika kunnossa, ainakin urologian osalta” (I4) **(palaute → virheettömyys + uskottavuus)**.

Tiedon hyödyntämisessä on yhtenä viestinnällisenä haasteena se, että ”kaikki, jotka vois sitä tiedolla johtamista tehdä, ei ehkä ihan vielä tiedä, että semmonen mahdollisuus on” (I3) **(viestintä → lisäarvo).**

6 TULOSTEN TARKASTELU JA POHDINTA: KLIINISEN TIETOALLASTIEDON LAADUN KEHITTÄMINEN VAATII ERILAISIA RESURSSEJA

Tässä luvussa vastataan tutkimuskysymyksiin edellisessä luvussa esiteltyjen tulosten pohjalta niitä aiempaan tutkimuskirjallisuuteen peilaten. Lisäksi käsitellään tutkimuksen luotettavuutta, tieteellistä, yhteiskunnallista ja käytännön merkitystä sekä jatkotutkimusaiheita.

6.1 Tulosten tarkastelu

Tapaustutkimuksen tarkoituksena oli selvittää sosioteknisestä ja kontekstuaalisesta tiedon hyödyntämisen näkökulmasta, mitä tiedon laatuongelmia suomalaisessa terveydenhuollon big datassa on, ja mistä ne aiheutuvat. Tutkimuksessa tarkasteltiin tietoallastiedon laatuongelmia ja niiden syitä tutkimuksessa ja tiedolla johtamisessa VSSHP:ssa ja urologian klinikalla sekä sitä, miten urologian tietoallashankkeessa on pyritty näitä ongelmia ratkaisemaan.

Kuviot 5 ja 6 tiivistävät vastauksen ensimmäiseen tutkimuskysymykseen ketjuiksi toisiinsa vaikuttavia laatuongelmien syitä ja laatuongelmia lähtien tiedon kirjaamisesta tietojärjestelmiin ja päätyen sen hyödyntämisen ongelmiin. Tulokset ovat yhdensuuntaiset Strongin ym. (1997) tutkimuksen kanssa, jonka mukaan tiedon laatuongelmat syntyvät monivaiheisessa sosioteknisessä prosessissa, jossa myös tiedon eri laatuominaisuudet ovat keskenään syy-seuraussuhteissa.

Analyysissä käytetyn mukaillun Wangin ja Strongin (1996) tiedon laatuulottuvuuksien viitekehyksen (taulukko 1) keskeisimpinä tiedon laatuominaisuuksina tapauskontekstissa näyttäytyivät tiedon *yhdenmukaisuus, teknisen käsittelyn helppous, saatavuus, täydellisyys, virheettömyys, ajantasaisuus* sekä *relevanssi* ja *lisäarvo*. Lisäksi analyysissä aineistosta nousivat viitekehystä täydentäviksi laatuominaisuuksiksi *rakenteisuus, tarkkuus,*

vertailtavuus, yhdisteltävyys, yhtenäisyys ja *eheys*. Nämä laatuominaisuudet selittävät eksplisiittisesti muiden laatuominaisuuksien välisiä yhteyksiä ja syy-seuraussuhteita sekä tuovat esiin ja perustelevat laatuominaisuuksien päällekkäisyyksiä (ks. Warwick ym., 2015) tapauskontekstissa. Ne ovat myös pääosin yhteneväisiä terveydenhuollon big datan ja sairauskertomustiedon laatuongelmia koskevan aiemman tutkimuksen kanssa (ks. liite 1).

Tuloksia on mahdollista tarkastella jäävuorena, jonka huipulla tiedon käyttäjä arvioi aineiston käyttökelpoisuutta (*relevanssi*) ja sen käytön hyötyjä (*lisäarvo*) tutkimuksessa tai tiedolla johtamisessa. Tulosten mukaan relevanssiin ja lisäarvoon vaikuttavat tiedonkeruun ja tiedon jalostamisen suunnittelu- ja toteutustapa, osaamis- ja teknologiaresurssit, toimiva yhteistyö sekä tietoallasaineiston *täydellisyys, virheettömyys ja ajantasaisuus*. Weiskopfin ja Wengin (2013) mukaan nämä kolme ovat tärkeimmät sairauskertomustiedon laatuominaisuudet toissijaisessa käytössä. Tulosten perusteella voidaan päätellä, että myös VSSHP:ssä tutkija ja tiedolla johtaja arvioi aineiston käyttökelpoisuutta juuri näiden ominaisuuksien perusteella.

Kun tutkitaan tiedon laatuongelmien syntymisen säännönmukaisuuksia, ei voida jäädä tiedon laadunarvioinnin tasolle. On selvitettävä, mistä täydellisyyden, virheettömyyden ja ajantasaisuuden heikkoudet juontuvat. Suuri osa haastatelluista piti laatuongelmien tärkeimpänä juurisyynä *tietojen kirjaamisen tapaa*. Se on myös kirjallisuuden mukaan keskeisin sairauskertomustiedon epätäydellisyyden ja virheellisyyden lähde toissijaisessa käytössä. Usein ei kerätä sellaista tietoa, joka olisi hyödynnettävissä haluttuun tarkoitukseen. Koska tietojen kirjaaminen on pääosin manuaalista, tallentamisessa tapahtuu inhimillisiä virheitä. Lääkärit joutuvat työnsä kiireisessä arjessa jatkuvasti valitsemaan potilaan hoidon ja kirjaamisen välillä. Potilastietojärjestelmät eivät toimi yhteen, niiden käyttö on hankalaa, kirjaamisperusteet ovat ristiriitaisia ja lääkäreiden motivaatiossa ja tietoteknisissä taidoissa on puutteita. Näiden syiden lisäksi tapaustutkimuksen tuloksissa näkyi, että kirjaamismotivaatio on alhainen, koska lääkärit eivät koe kirjaamisesta olevan hyötyä omassa työssään.

Kirjatussa tiedossa on siis paljon puutteita ja virheitä. Kirjallisuuden mukaan kirjatun tiedon *rakenteettomuus* ja *epäyhdenmukaisuus* ovat kuitenkin tärkeimmät sairauskertomustiedon toissijaisen hyödyntämisen ja terveydenhuollon big datan hyödyntämisen haasteet. Tutkimuksen tulokset ovat samankaltaiset. Usein lääkäri kirjaa potilaan tiedot narratiivisena tekstinä rakenteettomassa muodossa, jota on paljon ja se on epäyhdenmukaista. Epäyhdenmukaisuutta seuraa myös erilaisista tallentamisen tavoista ja muodoista, jotka vaihtelevat ajan, paikan, kirjaajan, ohjeistuksen, koodistojen ja tietojärjestelmän mukaan.

Koska terveydenhuollon tieto on erityisen moninaista ja sitä yhdistellään lukuisista eri lähteistä, epäyhdenmukaisuudesta aiheutuvat ongelmat kertautuvat. Yksi vaikeimmista ongelmista liittyy tietorakenteiden ja -tyyppien yhdistämiseen. Tämän tutkimuksen tulosten mukaan viiteavainten puute (alhainen *yhdisteltävyys*) ja tietojen suuri pirstaleisuus (alhainen *yhtenäisyys*) lisää virheiden mahdollisuutta jo organisaation sisäisiä tietolähteitä käytettäessä ja hankaloittaa tiedon teknistä käsittelyä. Monia vaiheita ja käsittelijöitä sisältävän jalostusprosessin epävakaus oli toinen keskeinen virhelähde. Jalostusvaiheessa eniten täydellisyyden, virheettömyyden ja ajantasaisuuden ongelmia aiheutui tietoallaslatausten epäonnistumisesta ja latausten erilaisista ajastuksista.

Tulosten mukaan tiedon saatavuuden ongelmat voivat johtaa tiedon epätäydellisyyteen toisiokäytössä. Strong ym. (1997) osoittivat, että tiedon käyttäjän näkökulmasta saatavuus ei liity vain tekniseen tietoon pääsyyn vaan myös tiedon käytön helppouteen ja nopeuteen. Tulosten mukaan rakenteettoman ja epäyhtenäisen tiedon saatavuus oli heikko, koska sen käyttö oli hankalaa, kallista tai mahdotonta. Big data -analytiikan tutkimus korostaa tiedon teknisen käsittelyn kustannustehokkuutta ja helppoutta arvon luomisessa (esim. Merino ym., 2015). Nämä ominaisuudet olivat odotetusti tiedon saatavuuteen vaikuttavia tekijöitä myös tutkimuskontekstissa. Tulosten perusteella teknisen käsittelyn helppouden ja kustannustehokkuuden lisääminen saatavuuden laadun ulottuvuuteen tarkentaisi ymmärrystä tiedon hyödyntämisen ongelmista terveydenhuollon tiedon toissijaisessa käytössä ja big datan jalostuksessa.

Saatavuutta hankaloittaa myös terveydenhuollon tietojen vahva tietosuoja ja sijainti fyysisesti ja teknisesti hajallaan eri tietorakenteissa ja rekisterinpitäjien tietovarastoissa. Tietoallas on jo sinänsä parantanut ja luonut toisiokäytön mahdollisuuksia VSSHP:ssa ja sen ulkopuolella. Tietoaltaassa on kuitenkin toistaiseksi vasta vähän perusterveydenhuollon tietoa. Esimerkiksi urologian klinikan potilaan hoitopolku sisältää vain erikoissairaanhoidon osuuden ja on siksi potilaan näkökulmasta puutteellinen.

Tietosuojaan ja -turvaan liittyvät tiedon saatavuusongelmien syyt eivät korostuneet haastatteluissa, joissa kyse oli organisaation sisäisestä tiedon käytöstä. Saman tiedon saatavuus tutkimukseen on kuitenkin alhaisempi kuin sen saatavuus tiedolla johtamiseen, koska tiedolla johtaja ei tarvitse erillistä lupaa tietojen käyttöön. VSSHP:ssä korkeatasoinen tekninen osaaminen ja teknologiaresurssit, kuten parempi laskentateho ja uudet tiedon louhimisen menetelmät paransivat tietoaltaan rakenteettoman tiedon saatavuutta.

Tiedon ymmärrettävyyteen ja jäljitettävyyteen liittyvät tekijät ovat tärkeitä toissijaisessa käytössä. Kliinikon substanssiosaamisen merkitys tietojärjestelmien ja kirjaamistapojen tuntijoina ja toisaalta lääketieteen asiantuntijoina korostui tuloksissa. Kliinikon tietotaidon kerrottiin olevan tärkeää, jotta osataan valita relevantteja tietoja ja tehdä oikeita johtopäätöksiä. Myös teknisestä näkökulmasta tiedon merkitykseen, alkuperään ja muokkaushistoriaan liittyvä tieto oli tärkeää. Toistaiseksi tietoallasympäristön tiedonhallinta vaati paljon manuaalista työtä, ja tiedon virheettömyyttä oli prosessissa mahdotonta täysin varmistaa. Ajantasaista ja riittävää metatietoa tietoallaslatausten sisältämästä tiedosta ei usein ollut saatavilla järjestelmätoimittajien salassapitosäädösten vuoksi. Tapauskontekstissa tiedon jalostusprosessin eri osapuolten sujuva yhteistyö, viestintä ja palautteenanto olikin erittäin tärkeää tiedon laadunhallinnan kannalta. Organisaation oman tietopalvelun sijainti fyysisesti sairaalan yhteydessä oli keskeinen mahdollistaja tiiviille yhteistyölle.

Urologian klinikalla koetut tiedon laatuongelmat olivat samankaltaisia kuin edellä on kuvattu. Tietoaltaan tieto ei ollut relevanttia eikä jalostettavissa eturauhassyövän hoidon laadun mittaamiseen. Ratkaisuksi klinikalle kehitettiin oma laaturekisterisovellus, jonka avulla urologit keräävät potilas- ja hoitotietoa eturauhassyöpäpotilaiden käynneillä. Käytetyt mittarit ovat rakenteisia ja ne on muotoiltu kansallisesti urologien yhteistyönä ja läheisessä yhteydessä THL:n eturauhassyövän laaturekisterihankkeeseen. Näin tavoitellaan kansallisesti yhdenmukaista, saatavuudeltaan hyvää, relevanttia tietoa, jolla on suuri

lisäarvo. Koska urologian laaturekisteritietoa on kerättävä kaksoiskirjaamalla, riskinä on, että lääkärit eivät jaksa kirjata tietoja, jolloin tietojen täydellisyys kärsii. Kliinikoiden kirjaamismotivaatiota saattaa kuitenkin parantaa mahdollisuus saada vertailutietoa leikkaustensa onnistumisesta ja näin oppia työstään.

Urologian klinikan tietoallashanke alkoi alun perin potilastietojen visuaalisen aikajananäkymän kehittämisestä. Näkymä tarjoaa lääkärille potilaan tiedot ruudulle tiiviissä muodossa tietoaltaan tietoja jalostamalla. Lisäksi on aggregoituja näkymiä tiedolla johtamisen tarpeisiin. Vaikka näkymän prosessi on tuotteistettu, se ei ole riittävän luotettava, jotta näkymä voisi saada lääkinnällisen laitteen aseman, ja sitä voitaisiin käyttää potilaan hoitoon. Kuten edellä on tullut ilmi, tietoallastiedon hyödyntämisprosessi on virhealtis. Tietoallaslatauksilla on virhelähteitä, joita on vaikea kontrolloida. Laatuongelmia aiheutuu lisäksi tiedon monista siirto-, lataus- ja muunnosvaiheista. Näiden seurauksena näkymä voi antaa vanhaa tai puutteellista tietoa potilaasta ja hänen hoidostaan, mikä saattaa vaikuttaa kliinikon tekemien laatukirjausten laatuun.

Tutkimus vahvisti ja täydensi aiempaa tietoa sairauskertomustiedon laatuongelmista ja niiden syistä toissijaisessa käytössä. Lisäksi se tuotti uutta, laadullista tietoa siitä, millaisen prosessin seurauksena tiedon toissijaisen käytön laatuongelmat syntyvät suomalaisessa terveydenhuollon tietoallasympäristössä. Keskeinen laatuongelmien syy oli hoidon yhteydessä kirjatun tiedon alhainen laatu toissijaisen käytön kontekstissa. Lisäksi lisäarvon jalostaminen potilastiedoista estyy, jos organisaatiolla ei ole siihen käytettävissä kliinikoiden ja it-asiantuntijoiden korkeatasoista osaamista tai kehittynyttä teknologiaa. Yksin osaamis- ja teknologiaresurssit eivät riitä, vaan kehittämistyössä tarvitaan myös pitkäjänteistä ja tiivistä lääketieteen, analytiikan ja informaatioteknologian asiantuntijoiden yhteistyötä.

6.2 Tutkimuksen luotettavuus

Tutkimuksen luotettavuuden arvioimiseksi tutkimusprosessi ja tutkijan valinnat on pyritty kuvaamaan riittävällä tarkkuudella. Tapauksen ja sen kontekstin kuvaukseen sekä koodaus- ja luokitteluprosessin läpinäkyvyyteen on kiinnitetty erityistä huomiota.

Aineiston ja haastateltavien triangulaatiolla vahvistettiin tutkimuksen uskottavuutta. On kuitenkin huomattava, että viralliset dokumentit urologian tietoallashankkeesta puuttuivat aineistosta eikä tutkija haastatellut kliinikoiden ja it-asiantuntijoiden lisäksi muita ammattiryhmiä. Esimerkiksi sairaanhoitajien ja taloushallinnon henkilöstön haastatteluilla olisi mahdollisesti saatu uutta tietoa. It-palveluntarjoajalta ei saatu haastatteluun tietoaltaan teknisistä työvaiheista käytännön vastuussa ollutta henkilöä.

Tutkija pyrki varautumaan tutkimukseen vaikuttaviin tekijöihin hankkimalla esitietoa ennen haastatteluvaihetta. Esitiedon puute hankkeesta ja haastateltavien osuudesta siinä aiheutti kuitenkin jonkin verran epävarmuutta tutkimusprosessiin. Tapauksen rajaaminen uudelleen aineistonkeruun alussa saattoi vaikuttaa haastatteluaineiston laatuun, vaikka haastattelurunko oli joustava. Kysymysmuotoiluiden ja haastateltavien kokemustaustan vuoksi aineistoa kertyi enemmän tietojen kirjaamiseen kuin tekniseen prosessiin

liittyvistä tiedon laatuongelmista. Lääkärien haastatteluissa urologien näkemykset korostuivat.

Tutkimuksen keskeinen heikkous on, että tulokset perustuvat vain yhteen tapaukseen, pieneen aineistoon sekä yhden tutkijan työhön ja subjektiivisiin tulkintoihin. Tapauksen uskottavuuden ja siirrettävyyden kannalta on tärkeää, että tapaus ja sen konteksti voitiin esitellä nimellä. Haastateltavien vastauksiin on kuitenkin saattanut vaikuttaa halu säilyttää myönteinen julkikuva ja se, että heidän anonymiteettiään ei voitu varmistaa.

6.3 Tutkimuksen tieteellinen, yhteiskunnallinen ja käytännöllinen merkitys

Tutkimuksella on tieteellistä merkitystä osana big datan tutkimusta terveydenhuollossa. Koska empiirinen tutkimus big datan laadusta ja big datasta sosioteknisestä näkökulmasta on ollut niukkaa, tutkimus on tieteellisesti tärkeä avaus. Terveydenhuolto on otollinen ympäristö tällaiselle big datan laadun tutkimukselle, koska terveydenhuollossa tiedon laatu on sekä erityisen kriittistä että siinä on paljon ongelmia. Ongelmat ovat esimerkiksi sairauskertomustiedon kirjaamiseen, tiedon moninaisiin lähteisiin, rakenteisiin ja tyyppeihin, keruu- ja käyttökonteksteihin sekä tiedonkulkuun liittyviä. Lisäksi big data -ympäristöjen tiedon laadunhallinta terveydenhuollossa on haasteellista.

Koska sairauskertomus on keskeisin big datan lähde terveydenhuollossa, terveydenhuollon tiedon ja samalla big datan laatuongelmat ovat usein sairauskertomustiedon laatuongelmia. Tutkimuksessa tuotettiin uutta tietoa tiedon laatuongelmien syntyprosesseista sairauskertomustiedon toisiokäytössä organisaation sisällä. Tutkimus paitsi vahvisti ja täydensi aiempien tutkimusten tuloksia, myös tarjosi holistisen tapauskuvauksen laatuongelmien syntyprosessista tapauskontekstissa. Tiettävästi vastaavaa ei ole aiemmin tehty. Lisäksi tapauskontekstissa hyvin moninaista, laadultaan vaihtelevaa ja jokseenkin runsasta tietoa käsiteltiin tietoallasympäristössä. Kontekstissa korostuivat monet big datalle ominaiset tiedon laadun haasteet.

Tutkimuksessa sovellettiin ja täydennettiin Wangin ja Strongin (1996) tiedon laatu-ulottuvuuksien viitekehystä. Tiedon laatuongelmien sosioteknisten syiden luokittelu tehtiin aineistolähtöisesti. Sekä täydennettyä laatu-ulottuvuuksien viitekehystä että laatuongelmien syiden luokittelua voidaan jatkossa soveltaa erilaisissa empiirisissä tutkimuksissa, joissa tutkitaan terveydenhuollon tiedon tai big datan laatua.

Terveydenhuollon tiedon laatu on myös yhteiskunnallisesti merkittävä tutkimusaihe. Toimiva ja kustannustehokas terveydenhuolto on yhteiskunnan tärkeimpiä voimavaroja. Big datan hyödyntämistä pidetään keinona tarjota laadukkaita terveyspalveluja väestön ikääntyessä ja resurssien niukentuessa. Tällä hetkellä terveydenhuollon toiminnassa kerätyn tiedon laatu on huono, ja sen hyödyntäminen organisaation sisälläkin hankalaa. Sairauskertomustiedon laatuongelmat on ratkaistava ennen kuin big datan hyödyntäminen laajemmin turvallisuuskriittisessä potilaiden hoidossa on mahdollista. Toissijaisessakin käytössä tiedon laatuongelmat aiheuttavat kustannuksia, ja myös vakavammat seuraukset ovat mahdollisia.

Tutkimus on terveydenhuollon tiedon toisiokäytön osalta hyvin ajankohtainen. Tutkimuksen tulokset voivat auttaa terveydenhuollon asiantuntijoita ymmärtämään tiedon

laatua, siihen vaikuttavia tekijöitä ja laatuongelmien syntymistä laajasta näkökulmasta. Ne voivat auttaa edistämään terveydenhuollon toiminnassa kerättävän tiedon laatua ja hyödyntämistä. Tulokset ovat kontekstisidonnaisia, joten erityistä hyötyä niistä voi olla VSSHP:n organisaatiossa. Tutkimuksesta voivat hyötyä myös muut organisaatiot, joissa kerätään, jalostetaan ja käytetään moninaista laatu- ja tietoturvakriittistä tietoa big data -ympäristössä.

6.4 Jatkotutkimusaiheita

Big datan laatuongelmien esille tuominen on paitsi tietojärjestelmätieteilijöiden moraalinen velvollisuus (Clarke, 2016) myös edellytys sille, että terveydenhuollon big datalle asetetut suuret odotukset voidaan lunastaa. Olisi tärkeää tutkia laajasti terveystiedon hyödyntämisen esteitä sen toissijaisissa käyttötarkoituksissa, kuten tutkimuksessa, tiedolla johtamisessa ja innovaatiotoiminnassa. Tiedon laatuongelmia on tutkittava sellaisista näkökulmista ja sellaisin lähestymistavoin, jotka ottavat huomioon organisaatiotekijät, ihmiset, prosessit ja teknologian, tavoitteet ja arvot sekä laajemman toimintaympäristön. Tiedon tekninen laadunarviointi ei riitä, vaan tietoa tulee tutkia sosioteknisenä ja yhteiskunnallisena ilmiönä. Terveydenhuollossa tietoon liittyy poliittisia ja taloudellisia intressejä. Kun tiedosta jalostetaan arvoa, olisikin keskeistä tutkia, mitkä tavoitteet ja arvot käytännössä ohjaavat tiedon keräämistä, jalostamista ja käyttöä. Eräs tähän liittyvä tutkimuskysymys voisi olla, miten tulosohjaus vaikuttaa terveydenhuollon tulosmittareista tuotettuun tietoon. Lisäksi tarvitaan empiiristä tutkimusta myös terveydenhuollon tiedon ja big datan teknisestä laadusta.

Tutkimusta voisi luontevasti jatkaa tutkimalla terveydenhuollon big datan laatua erilaisissa konteksteissa alkaen siitä, millaisia tiedon laatuongelmia VSSHP:n tietoallasta hyödyntävät ulkopuoliset tutkijat ja hankkeet ovat kohdanneet. Muita mahdollisia tutkimuskohteita ovat esimerkiksi THL:n laaturekisterihankkeet sekä Findatan ja Kelan tietoasiakkaat ja kansainvälinen yhteistyö. Myös terveydenhuollon tiedon käyttö tiedolla johtamisessa päätöksenteon eri tasoilla on ajankohtainen tutkimusaihe. Paitsi toissijaisessa käytössä, terveystiedon laatua ja big datan käytön mahdollisuuksia olisi tutkittava potilaiden hoidossa, jossa tiedon laatu vaikuttaa kriittisesti potilasturvallisuuteen. Toimintaympäristön kannalta keskeisiä tutkimuksen aihepiirejä ovat yhteiskunnan luomat edellytykset tiedon toisiokäytölle, kuten toisiolaki ja sen toimeenpano, tutkimuksen ja innovaatiotoiminnan rahoitus ja muu terveystiedon toisiokäytön resursointi.

LÄHTEET

Abbasi, A., Sarker, S., & Chiang, R. (2016). Big Data Research in Information Systems: Toward an Inclusive Research Agenda. *Journal of the Association for Information Systems*, *17*(2), I–XXXII.

Ackoff, R. (1989). From Data to Wisdom. *Journal of Applied Systems Analysis*, *16*(1989), 3–9.

Alastalo, M., Vaittinen, T., & Åkerman, M. (2017). Asiantuntijahaastattelu. Teoksessa M. Hyvärinen, P. Nikander & J. Ruusuvuori (toim.), *Tutkimushaastattelun käsikirja* (211-230). Tampere: Vastapaino.

Ardagna, D., Cappiello, C., Samá, W., & Vitali, M. (2018). Context-aware data quality assessment for big data. *Future Generation Computer Systems*, *89*, 548–562.

Brynjolfsson, E. and McAfee, A. (2014). *The Second Machine Age: Work, progress, and prosperity in a time of brilliant technologies.* New York: WW Norton & Company.

Auffray, C., Balling, R., Barroso, I., Bencze, L., Benson, M., Bergeron, J., … Zanetti, G. (2016). Making sense of big data in health research: Towards an EU action plan. *Genome Medicine*, *8*(1), 1–13.

Auria Tietopalvelu (2020a). *Aineistot*. Haettu 6.11.2020 osoitteesta https://www.auria.fi/tietopalvelu/tutkijalle/index.html#aineistot

Auria Tietopalvelu (2020b). Tehtävämme. Haettu 6.11.2020 osoitteesta https://www.auria.fi/tietopalvelu/index.html#tehtavat

Baesens, B., Bapna, R., Marsden, J. R., Vanthienen, J., & Zhao, J. L. (2016). Transformational Issues of Big Data and Analytics in Networkded Business. *MIS Quarterly*, *40*(4), 807–818.

Bai, L., Meredith, R., & Burstein, F. (2018). A data quality framework, method and tools for managing data quality in a health care setting: an action case study. *Journal of Decision Systems*, *27*, 144–154.

Baro, E., Degoul, S., Beuscart, R., & Chazard, E. (2015). Toward a literature-driven definition of big data in healthcare. *BioMed Research International*.

Barriball, K. L., & While, A. (1994). Collecting data using a semi-structured interview: a discussion paper. *Journal of Advanced Nursing*, *19*, 328–335.

Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). Big Data In Health Care: Using Analytics To Identify And Manage High-Risk And High-Cost Patients. *Health Affairs*, *33*(7), 1123–1131.

Batini, C., Palmonari, M., & Viscusi, G. (2014). Opening the Closed World: A Survey of Information Quality Research in the Wild. Teoksessa L. Floridi & P. Illari (toim.), *The Philosophy of Information Quality* (43–73). Cham: Springer.

Bayley, K. B., Belnap, T., Savitz, L., Masica, A. L., Shah, N., & Fleming, N. S. (2013). Challenges in using electronic health record data for CER: Experience of 4 learning organizations and solutions applied. *Medical Care*,

51(8 SUPPL.3), 80–86.

Beyer M. A., Laney D. (2012) *The importance of 'big data': a definition*. Stamford: Gartner.

Bogner, A., Littig, B. & Menz, W. (2009): Introduction: Expert Interviews – An Introduction to a New Methodological Depate. Teoksessa A. Bogner, B. Littig & W. Menz (toim.), *Interviewing Experts. Research Methods Series. (1-16)*. Lontoo: Palgrave Macmillan.

Bonimi, S. (2016). The Electronic Health Record: A Comparison of Some European Countries. Teoksessa *Information and Communication Technologies in Organizations and Society (15),* 33–50.

Borycki, E. (2013). Trends in health information technology safety: From technology-induced errors to current approaches for ensuring technology safety. *Healthcare Informatics Research*. Korean Society of Medical Informatics.

Botsis T., Hartvigsen G., Chen F., and Weng C. (2010). Secondary use of EHR: Data Quality Issues and Informatics Opportunities. *Summit on Translational Bioinformatics 2010 (2010),* 1-5.

Buhl, H. U., Röglinger, M., Moser, F., & Heidemann, J. (2013). Big Data. A Fashionable Topic with(out) Sustainable Relevance for Research and Practice? *Business & Information Systems Engineering*, *5*(2), 65–69.

Cai, L., & Zhu, Y. (2015). The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal*, *14*(0), 2.

Cappiello, C., Samá, W., & Vitali, M. (2018). Quality awareness for a successful big data exploitation. Teoksessa *ACM International Conference Proceeding Series (37–44).* New York: Association for Computing Machinery.

Carter, S. M., & Little, M. (2007). Taking Action : Epistemologies , Methodologies , and Methods in Qualitative Research. *Qualitative Health Research*, *17*(10), 1316–1328.

Chae, B. (2019). A General framework for studying the evolution of the digital innovation ecosystem: The case of big data. *International Journal of Information Management*, *45*, 83–94.

Chan, K. S., Fowles, J. B., & Weiner, J. P. (2010). Electronic Health Records and the Reliability and Validity of Quality Measures: A Review of the Literature. *Medical Care Research and Review*, *67*(5), 503–527.

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*, *36*(4), 1165-1188.

Chen, M. C., Ball, R. L., Yang, L., Moradzadeh, N., Chapman, B. E., Larson, D. B., … Lungren, M. P. (2018). Deep learning to classify radiology Free-Text reports. *Radiology*, *286*(3), 845–852.

Cichy, C., & Rass, S. (2019). An overview of data quality frameworks. *IEEE Access*, *7*, 24634–24648.

Cirillo, D., & Valencia, A. (2019). Big data analytics for personalized medicine. *Current Opinion in Biotechnology*, *58*, 161–167.

Clarke, R. (2016). Big data, big risks. *Information Systems Journal*, *26*(1), 77–90.

Constantiou, I. D., & Kallinikos, J. (2015). New games, new rules: Big data and

the changing context of strategy. *Journal of Information Technology*, *30*(1), 44–57.

Costa, F. F. (2014). Big data in biomedicine. *Drug Discovery Today*, *19*(4), 433–440.

Cridland, E. K., Jones, S. C., Caputi, P., & Magee, C. A. (2015). Qualitative research with families living with autism spectrum disorder: Recommendations for conducting semistructured interviews. *Journal of Intellectual and Developmental Disability*, *40*(1), 78–91.

Darke, P., Shanks, G., & Broadbent, M. (1998). Successfully completing case study research: combining rigour, relevance and pragmatism. *Information Systems Journal*, *8*(4), 273–289.

Darst, R., Hakala, M., Kaski, K. (2018). Isaacus-hankkeen tietoallasratkaisujen arviointi tutkimuskäytössä. Tietotekniikan laitos. Aalto yliopisto. Noudettu 6.11.2020 osoitteesta https://media.sitra.fi/2017/02/06161409/tietoaltaanarviointi30042017final.pdf

Davenport, T. H. (2018). From analytics to artificial intelligence. *Journal of Business Analytics*, *1*(2), 73–80.

Davenport, T. H., Barth, P. F. P., & Bean, R. (2012). How ' Big Data ' is Different. *MITSloan Management Review, 54* (1), 21–24.

Davis, C. K. (2014). Beyond data and analysis. *Communications of the ACM*, *57*(6), 39–41.

De Mauro, A., Greco, M., & Grimaldi, M. (2016). A formal definition of Big Data based on its essential features. *Library Review*, *65*(3), 122–135.

DeLone, W. H., & McLean, E. R. (1992). Information Systems Success: The Quest for the Dependent Variable. *Information Systems Research*, *3*(1), 60–95.

DeLone, W. H., & McLean, E. R. (2003). The DeLone and McLean Model ofInformation Systems Success: A Ten-Year Update. *Journal of Management Information Systems*, *19*(4), 9–30.

Demchenko, Y., Grosso, P., de Laat, C., & Membrey, P. (2013). Addressing big data issues in Scientific Data Infrastructure. Teoksessa *2013 International Conference on Collaboration Technologies and Systems (CTS) (48–55)*. San Diego, California, May 20-24, 2013.

Demirkan, H., & Delen, D. (2013). Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud. *Decision Support Systems*, *55*(1), 412–421.

Dentler, K., Cornet, R., Teije, A. ten, Tanis, P., Klinkenbijl, J., Tytgat, K., & Keizer, N. de (2014). Influence of data quality on computed Dutch hospital quality indicators: A case study in colorectal cancer surgery. *BMC Medical Informatics and Decision Making*, *14*(1).

Dinov, I. D. (2016). Volume and value of big healthcare data. *Journal of Medical Statistics and Informatics*, *4*(3), 1–7.

Dungey, S., Glew, S., Heyes, B., Macleod, J., & Tate, A. R. (2016). Exploring practical approaches to maximising data quality in electronic healthcare records in the primary care setting and associated benefits. Report of panel-led discussion held at SAPC in July 2014. *Primary Health Care Research and*

Development, *17*(5), 448–452.

Ehrenstein, Vera;, Nielsen, H., Pedersen, A. B., Johnsen, S. P., & Pedersen, L. (2017). Clinical epidemiology in the era of big data: new opportunities, familiar challenges. *Clin Epidemiol., 9*, 245–250.

Elo, S., Kääriäinen, M., Kanste, O., Pölkki, T., Utriainen, K., & Kyngäs, H. (2014). Qualitative Content Analysis. *SAGE Open*, *4*(1), 1-10.

Eppler, M. J. (2001). The Concept of Information Quality: An Interdisciplinary Evaluation of Recent Information Quality Frameworks. *Studies in Communication Sciences, 1* (2), 167-182.

Eskola, J. & Suoranta, J. (1998). Johdatus laadulliseen tutkimukseen (3. painos). Tampere: Vastapaino.

Estabrooks, P. A., Boyle, M., Emmons, K. M., Glasgow, R. E., Hesse, B. W., Kaplan, R. M., … Taylor, M. V. (2012). Harmonized patient-reported data elements in the electronic health record: supporting meaningful use by primary care action on health behaviors and key psychosocial factors. *Journal of the American Medical Informatics Association*, *19*(4), 575–582.

Ettala, O., Boström, P., Santti, H., Kaipia, A., Järvinen, R., Nisen, H., Perttilä, I., Häkkinen, J., Raatikainen, S., Parpala, T., Matikainen, M. & Leppilahti, M. (2018). *Urologisten sairauksien laatu*. Suomen urologiyhdistys ry. Haettu osoitteesta https://www.urologiyhdistys.fi/@Bin/193264/Urologian+laatuka%CC%88sikirja.pdf

Feldman B, Martin E. M., Skotnes T (2012). *Big Data in Healthcare. Hype and Hope.* Dr. Bonnie 360 degree (Business Development for Digital Health). Haettu osoitteesta http://www.west-info.eu/files/big-data-in- healthcare.pdf

Galetsi, P., Katsaliaki, K., & Kumar, S. (2020). Big data analytics in health sector: Theoretical framework, techniques and prospects. *International Journal of Information Management*, *50*, 206–216.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144.

Gartner (2012). Gartner Glossary. Big data. Haettu 26.8.2020 osoitteesta https://www.gartner.com/en/information-technology/glossary/big-data

George, G., Osinga, E. C., Lavie, D., & Scott, B. A. (2016). Big Data and Data Science Methods for Management Research. *Academy of Management Journal*, *59*(5), 1493–1507.

Ghasemaghaei, M., & Calic, G. (2019). Can big data improve firm decision quality? The role of data quality and data diagnosticity. *Decision Support Systems*, *120*, 38–49.

Gill, P., Stewart, K., Treasure, E., & Chadwick, B. (2008). Methods of data collection in qualitative research: Interviews and focus groups. *British Dental Journal*, *204*(6), 291–295.

Goldberg, S. I., Niemierko, A., & Turchin, A. (2008). Analysis of data errors in clinical research databases. Teoksessa *AMIA Annual Symposium proceedings / AMIA Symposium*, *2008*, 242–246.

Günther, W. A., Rezazade Mehrizi, M. H., Huysman, M., & Feldberg, F. (2017). Debating big data: A literature review on realizing value from big data. *Journal of Strategic Information Systems*, *26*(3), 191–209.

Hammais, A., Varjonen, J. & Virkki A. (2018). *The Clinical Data Refinery. Management and Administration of the Analytics Environment.* Centre for Clinical Informatics, Hospital district of southwest Finland. Haettu osoitteesta https://github.com/Sitra-Isaacus/VSSHP-tietoallas-dap/blob/master/book/cci_book.pdf

Haseeb, A., & Pattun, G. (2017). A review on NoSQL: Applications and challenges. *International Journal of Advanced Research in Computer Science*, *8*(1), 203–207.

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Ullah Khan, S. (2015a). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems*, *47*, 98–115.

Hersh, W. R., Weiner, M. G., Embi, P. J., Logan, J. R., Payne, P. R. O., Bernstam, E. V., … Saltz, J. H. (2013). Caveats for the use of operational electronic health record data in comparative effectiveness research. *Medical Care*, *51*(8 SUPPL.3), 30–37.

Hirsjärvi, S., & Hurme, H. (2000). *Tutkimushaastattelu. Teemahaastattelun teoria ja käytäntö.* Helsinki: Yliopistopaino.

Hoepfl, M. C. (1997). A Primer for Technology Education Researchers. *Journal of Technology Education*, *9*(1), 47–63.

Hoffman, S. (2014). Medical Big Data and Big Data Quality Problems. *SSRN Electronic Journal*.

Hoffman, S., & Podgurski, A. (2013). The use and misuse of biomedical data: is bigger really better? *American journal of law & medicine*, *39*(4), 497–538.

Holve, E., Kahn, M., Nahm, M., Ryan, P., & Weiskopf, N. (2013). A comprehensive framework for data quality assessment in CER. Teoksessa *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, *2013*, 86–88.

Hripcsak, G., Knirsch, C., Zhou, L., Wilcox, A., & Melton, G. (2011). Bias Associated with Mining Electronic Health Records. *Journal of Biomedical Discovery and Collaboration*, *6*, 48–52.

Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, *15*(9), 1277–1288.

Hyppönen, H, Winblad, I., Reponen, J., Lääveri, T., & Vänskä, J. (2012). *Lääkärien kokemukset alueellisesta potilastiedon vaihdosta.* Terveyden ja hyvinvoinnin laitos. Haettu osoitteesta http://www.thl.fi/thl-client/pdfs/f4191f01-b6f7-46c0-b0eb-8358a66aca39

Hyppönen, H. & Ilmarinen, K. (2016). *Sosiaali- ja terveydenhuollon digitalisaatio.* Tutkimuksesta tiiviisti 22/2016. Helsinki: Terveyden ja hyvinvoinnin laitos

Hyppönen, H., Lääveri, T., Hahtela, N., Suutarla, A., Sillanpää, K., Kinnunen, U.-M., … Saranto, K. (2018). Kyvykkäille käyttäjille fiksut järjestelmät? Sairaanhoitajien arviot potilastietojärjestelmistä 2017. *Finnish Journal of eHealth and eWelfare*, *10*(1), 30–59.

ISO = International Standardization Organization (2008). *ISO/IEC 25012:2008*

Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model. Haettu 26.8.2020 osoitteesta https://www.iso.org/standard/35736.html.

Jalonen, H. (2015). Tiedolla johtamisen näyttämö ja kulissit. Teoksessa Virtanen, P., Stenvall, J. & Rannisto P.-H., *Tiedolla johtaminen - teoraa ja käytäntöjä* (40–68). Tampere: Tampereen yliopistopaino Oy - Juvenes Print.

Jee, K., & Kim, G.-H. (2013). Potentiality of Big Data in the Medical Sector: Focus on How to Reshape the Healthcare System. *Healthcare Informatics Research*, *19*(2), 79–85.

Jetley, G., & Zhang, H. (2019). Electronic health records in IS research: Quality issues, essential thresholds and remedial actions. *Decision Support Systems*, *126*, 113–137.

Johnson, S. G., Speedie, S., Simon, G., Kumar, V., & Westra, B. L. (2015). A Data Quality Ontology for the Secondary Use of EHR Data. Teoksessa *AMIA Annual Symposium proceedings. AMIA Symposium*, *2015* (1937–1946). Haettu osoitteesta https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4765682/

Jones, M. (2019a). What we talk about when we talk about (big) data. *The Journal of Strategic Information Systems*, *28*(1), 3–16.

Kahn, B. K., Strong, D. M., & Wang, R. Y. (1997). A Model for Delivering Quality Information as Product and Service. Teoksessa *Conference on Information Quality* (80–94). Cambridge, MA, 1997.

Kahn, Beverly K., Strong, D. M., & Wang, R. Y. (2002). Information quality benchmarks. *Communications of the ACM*, *45*(4), 184-192.

Kahn, M. G., Callahan, T. J., Barnard, J., Bauck, A. E., Brown, J., Davidson, B. N., … Schilling, L. (2016). A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. *eGEMs (Generating Evidence & Methods to improve patient outcomes)*, *4*(1), 1-18.

Kallio, H., Pietilä, A., Johnson, M., & Kangasniemi, M. (2016). Systematic methodological review : developing a framework for a qualitative semi-structured interview guide. *Journal of Advanced Nursing, 72*(12), 2954–2965.

Kankanhalli, A., Hahn, J., Tan, S., & Gao, G. (2016). Big data and analytics in healthcare: Introduction to the special section. *Information Systems Frontiers*, *18*(2), 233–235.

Khine, P. P., & Wang, Z. S. (2018). Data lake: a new ideology in big data era. Teoksessa *ITM Web of Conferences*, *17* (1–10). Haettu osoitteesta: https://doi.org/10.1051/itmconf/20181703025

Kitchin, R. (2013). Big data and human geography: Opportunities, challenges and risks. *Dialogues in Human Geography*, *3*(3), 262–267.

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, 1, April–June 2014, 1–12.

Kitchin, R., & McArdle, G. (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society*, *3*(1), 1-10.

Kortekangas, P. (2016). TYKS tietoallaskonsepti. Haettu osoitteesta: www.healthhub.fi

Kruse, C. S., Goswamy, R., Raval, Y., & Marawi, S. (2016a). Challenges and

Opportunities of Big Data in Health Care: A Systematic Review. *JMIR medical informatics*, *4*(4), e38, 1-12.

Kruse, C. S., Goswamy, R., Raval, Y., & Marawi, S. (2016b). Challenges and Opportunities of Big Data in Health Care: A Systematic Review. *JMIR Medical Informatics*, *4*(4), e38.

Laihonen, H., Hannula, M., Helander, N., Ilvonen, I., Jussila, J., Kukkola, M., … Yliniemi, T. (2013). *Tietojohtaminen.* Tampere: Tampereen teknillinen yliopisto, tiedonhallinnan ja logistiikan laitos.

Laine, S., Soikkeli, J., Ruohonen, T., & Nieminen, M. (2015). Timestamp Accuracy in Healthcare Business. Teoksessa *Proceedings of the 20th International Conference on Information Quality, ICIQ 2015* (150-164), Cambridge, Massachusetts, July 24, 2015.

Laney, D. (2001). *3-D data management: Controlling data volume, velocity and variety.* Application Delivery Strategies by META Group Inc.

Laranjeiro, N., Soydemir, S. N., & Bernardino, J. (2015). A Survey on Data Quality: Classifying Poor Data. Teoksessa *2015 IEEE 21st Pacific Rim International Symposium on Dependable Computing (PRDC)* (179–188), Zhangjiajie, 2015.

Latvala, E. & Vanhanen-Nuutinen, L. (2001). Laadullisen hoitotieteellisen tutkimuksen perusprosessi: Sisällönanalyysi. Teoksessa S. Janhonen & M. Nikkonen (toim.), *Laadulliset tutkimusmenetelmät hoitotieteessä (21-43).* Helsinki: Werner Söderström Osakeyhtiö.

Liaw, S.T., Rahimi, A., Ray, P., Taggart, J., Dennis, S., de Lusignan, S., … Talaei-Khoei, A. (2013). Towards an ontology for data quality in integrated chronic disease management: A realist review of the literature. *International Journal of Medical Informatics*, *82*(1), 10–24.

Liaw, S. T., Taggart, J., Dennis, S., & Yeo, A. (2011). Data quality and fitness for purpose of routinely collected data--a general practice case study from an electronic practice-based research network (ePBRN). Teoksessa *AMIA Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, *2011* (785–794).

Liyanage, H., de Lusignan, S., Liaw, S.-T., Kuziemsky, C., Mold, F., Krause, P., … Jones, S. (2014). Big Data Usage Patterns in the Health Care Domain: A Use Case Driven Approach Applied to the Assessment of Vaccination Benefits and Risks. *Yearbook of Medical Informatics*, *23*(01), 27–35.

Loshin, D. (2014). *Understanding Big Data Quality for Maximum Information Usability.* White Paper, SASA Institute Inc.

Lu, H. C., Hwang, F. J., & Huang, Y. H. (2020). Parallel and distributed architecture of genetic algorithm on Apache Hadoop and Spark. *Applied Soft Computing Journal*, *95*, 106497, 1-15.

Lääkäriliitto (2020). Potilasasiakirjat. Haettu 20.7.2020 osoitteesta: https://www.laakariliitto.fi/laakarinetiikka/potilas-laakarisuhde/potilasasiakirjat/

Malterud, K., Siersma, V. D., & Guassora, A. D. (2016). Sample Size in Qualitative Interview Studies: Guided by Information Power. *Qualitative*

Health Research, *26*(13), 1753–1760.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H., (2011) *Big data: The next frontier for innovation, competition, and productivity.* McKinsey Global Institute Haettu osoitteesta: https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation

Mehta, N., & Pandit, A. (2018). Concurrence of big data analytics and healthcare: A systematic review. *International Journal of Medical Informatics*, *114*(1), 57–65.

Merino, J., Caballero, I., Rivas, B., Serrano, M., & Piattini, M. (2016). A Data Quality in Use model for Big Data. *Future Generation Computer Systems*, *63*, 123–130.

Mikalef, P., Pappas, I. O., Krogstie, J., & Giannakos, M. (2018). Big data analytics capabilities: a systematic literature review and research agenda. *Information Systems and e-Business Management*, *16*(3), 547–578.

Mikalef, P., & Pateli, A. (2017). Information technology-enabled dynamic capabilities and their indirect effect on competitive performance: Findings from PLS-SEM and fsQCA. *Journal of Business Research*, *70*, 1–16.

Molinari, A., & Nollo, G. (2020). The quality concerns in health care Big Data, 302–305.

Murdoch, T. B., & Detsky, A. S. (2013). The Inevitable Application of Big Data to Health Care. *JAMA*, *309*(13), 1351.

Myers, M. D., & Newman, M. (2007). The qualitative interview in IS research: Examining the craft. *Information and Organization*, *17*(1), 2–26.

Mäkelä, K. (1990). Kvalitatiivisen analyysin arviointiperusteet. Teoksessa K. Mäkelä (toim.), *Kvalitatiivisen aineiston analyysi ja tulkinta (42–59)*. Helsinki: Gaudeamus.

Nelson, R. R., Todd, P. A., & Wixom, B. H. (2005). Antecedents of information and system quality: An empirical examination within the context of data warehousing. *Journal of Management Information Systems*, *21*(4), 199–235.

Newgard, C. D., Zive, D., Jui, J., Weathers, C., & Daya, M. (2012). Electronic versus manual data processing: Evaluating the use of electronic health records in out-of-hospital clinical research. *Academic Emergency Medicine*, *19*(2), 217–227.

Nobles, A. L., Vilankar, K., Wu, H., & Barnes, L. E. (2015). Evaluation of data quality of multisite electronic health record data for secondary analysis. *Proceedings - 2015 IEEE International Conference on Big Data, IEEE Big Data 2015*, 2612–2620.

Orfanidis, L., Bamidis, P. D., & Eaglestone, B. (2004). Health Informatics Journal, *10*(1), 23–36.

Patton, M. (2002). *Qualitative Research & Evaluation Methods* (3rd revised ed.). Thousand Oaks, CA: Sage Publications.

Pfadenhauer, M. (2009). At Eye Level: The Expert Interview — a Talk between Expert and Quasi-expert. Teoksessa A. Bogner, B. Littig, W. Menz (toim.), *Interviewing Experts (81-97)*. Research Methods Series. Lontoo: Palgrave Macmillan.

Pfeffer, J., Sutton, R.I. (2006). Evidence-based management. *Harvard Bus. Rev., 84* (1), 62.

Piri, S. (2020). Missing care: A framework to address the issue of frequent missing values. The case of a clinical decision support system for Parkinson's disease. *Decision Support Systems*, *136*(November 2019), 113339.

Porter, S. C., & Mandl, K. D. (1999). Data quality and the electronic medical record: a role for direct parental data entry. Teoksessa *Proceedings / AMIA Annual Symposium. AMIA Symposium (354–358).*

Price, R., & Shanks, G. (2004). A semiotic information quality framework. Teoksessa *Proceedings of the International Conference on Decision Support Systems DSS04 (658–672)*.

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*, *2*(1), 3.

Rahm, E., & Do, H. H. (2000). Data Engineering - Special Issue on Data Cleaning. *Data Engineering*, *23*(4), 3–13.

Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine*, *169*(12), 866–872.

Rao, D., Gudivada, V. N., & Raghavan, V. V. (2015). Data quality issues in big data. *Proceedings - 2015 IEEE International Conference on Big Data, IEEE Big Data 2015*, 2654–2660.

Ravat, F. & Zhao, Y. (2019). Data Lakes: Trends and Perspectives. *International Conference on Database and Expert Systems Applications (DEXA 2019),* Linz, Austria, 2019.

Rayner, J., Khan, T., Chan, C., & Wu, C. (2020). Illustrating the patient journey through the care continuum: Leveraging structured primary care electronic medical record (EMR) data in Ontario, Canada using chronic obstructive pulmonary disease as a case study. *International Journal of Medical Informatics*, *140*(April), 104159.

Reeves, C. A., & Bednar, D. A. (1994). Defining Quality: Alternatives and Implications. *Academy of Management Review*, *19*(3), 419–445.

Richesson, R. L., Horvath, M. M., & Rusincovitch, S. A. (2014). Clinical research informatics and electronic health record data. *Yearbook of medical informatics*, *9*(1), 215–223.

Rowley, J. (2007). The wisdom hierarchy: Representations of the DIKW hierarchy. *Journal of Information Science*, *33*(2), 163–180.

Ruusuvuori, J., Nikander, P. & Hyvärinen, M. (2010). Haastattelun analyysin vaiheet. Teoksessa J. Ruusuvuori, P. Nikander & M. Hyvärinen (toim.), *Haastattelun analyysi (9-36).* Tampere: Vastapaino

Sackett, D. L. (1996). Evidence based medicine: what it is and what it isn't. *BMJ, 312*, 71–72.

Sadiq, S., Yeganeh, N. K., & Indulska, M. (2011). 20 Years of Data Quality Research: Themes, Trends and Synergies. *Conferences in Research and Practice in Information Technology Series*, *115*, 153–162.

Sáez, C., Zurriaga, O., Pérez-Panadés, J., Melchor, I., Robles, M., & García-Gómez, J. M. (2016). Applying probabilistic temporal and multisite data

quality control methods to a public health mortality registry in Spain: A systematic approach to quality control of repositories. *Journal of the American Medical Informatics Association*, *23*(6), 1085–1095.

Sarafidis, M., Tarousi, M., Anastasiou, A., Pitoglou, S., Lampoukas, E., Spetsarias, A., … Koutsouris, D. (2020). Data quality challenges in a learning health system. *Studies in Health Technology and Informatics*, *270*, 143–147.

Sarajärvi, A., & Tuomi, J. (2018). *Laadullinen tutkimus ja sisällönanalyysi (uudistettu laitos).* Helsinki: Kustannusosakeyhtiö Tammi.

Sitra (2018). Isaacus-esituotantohankkeet. Haettu 6.11.2020 osoitteesta https://www.sitra.fi/hankkeet/isaacus-esituotantohankkeet/

Sitra (2019). *Kansallisen Kuva-mittariston pilotointi alueellisen tietoaltaan päällä. Loppuraportti. Versio 1.1.*

STM=Sosiaali- ja terveysministeriö (2019). *Laki sosiaali- ja terveystietojen toissijaisesta käytöstä (toisiolaki). Toimeenpanon valmistelun kokonaiskuva.* Sosiaali- ja terveysministeriön raportteja ja muistioita 2019:44, 2019.

STM (2020). Kanta-palvelujen kehittäminen. Noudettu 29.7.2020 osoitteesta: https://stm.fi/sotetiedonhallinta/kanta-palvelujen-kehittaminen

Stewart, R. (2002). *Evidence-based management: a practical guide for health professionals.* Oxon: Radcliffe Medical Press.

Strong, D. M., Lee, Y. W., & Wang, R. Y. (1997). Data quality in context. *Communications of the ACM*, *40*(5), 103–110.

Sukumar, S. R., Natarajan, R., & Ferrell, R. K. (2015). Quality of Big Data in healthcare. *International Journal of Health Care Quality Assurance*, *28*(6), 621–634.

Surbakti, F. P. S., Wang, W., Indulska, M., & Sadiq, S. (2020). Factors influencing effective use of big data: A research framework. *Information and Management*, *57*(1), 103146.

Szlezák, N., Evers, M., Wang, J., & Pérez, L. (2014). The Role of Big Data and Advanced Analytics in Drug Discovery, Development, and Commercialization. *Clinical Pharmacology & Therapeutics*, *95*(5), 492–495.

Taleb, I., Serhani, M. A., & Dssouli, R. (2018). Big Data Quality: A Survey. *Proceedings - 2018 IEEE International Congress on Big Data, BigData Congress 2018 - Part of the 2018 IEEE World Congress on Services*, 166–173.

Talha, M., El Kalam, A. A., & Elmarzouqi, N. (2019). Big data: Trade-off between data quality and data security. *Procedia Computer Science*, *151*, 916–922.

THL=Terveyden ja hyvinvoinnin laitos (2020). Tietolupaviranomainen Findata. Haettu 6.11.2020 osoitteesta https://thl.fi/fi/tilastot-ja-data/aineistot-ja-palvelut/tietolupaviranomainen-findata

UICC=The Union for International Cancer Control (UICC) (2020). What is TNM? Haettu 5.11.2020 osoitteesta https://www.uicc.org/resources/tnm

Valo, J. (2018). 2M-IT Data Lake - Success Story. Big Data Event. 15.5.2018. Haettu 5.11.2020 osoitteesta https://slideplayer.fi/slide/15145652/

Virkki, A. (2017). Tietoaltaan hyödyntäminen ja visualisoinnin tekniikat. Sosiaali- ja terveydenhuollon atk-päivät 23.–24.5.2017, Helsinki. Haettu 5.11.2020

osoitteesta http://atk-paivat.fi/2017/S04-Virkki.pdf

Wamba, S. F., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). How ”big data” can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics, 165*, 234–246.

Wamba, S. F., Akter, S., Trinchera, L., & De Bourmont, M. (2018). Turning information quality into firm performance in the big data economy. *Management Decision*, Online First 1-28.

Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, *39*, 86–95.

Wang, R. Y., & Strong, D. M. (1996a). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, *3*(4), 5–13.

Wang, R. Y., & Strong, D. M. (1996b). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, *12*(4), 5–33.

Wang, Y., Kung, L., Gupta, S., & Ozdemir, S. (2019). Leveraging Big Data Analytics to Improve Quality of Care in Healthcare Organizations: A Configurational Perspective. *British Journal of Management*, *30*(2), 362–388.

Warwick, W., Johnson, S., Bond, J., Fletcher, G., & Kanellakis, P. (2015). A Framework to Assess Healthcare Data Quality. *The European Journal of Social and Behavioural Sciences*, *13*(2), 1730–1735.

Weiskopf, N. G., & Weng, C. (2013). Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association*, *20*(1), 144–151.

Weiskopf, N. G., Hripcsak, G., Swaminathan, S., & Weng, C. (2013a). Defining and measuring completeness of electronic health records for secondary use. *Journal of Biomedical Informatics*, *46*(5), 830–836.

Weiskopf, N. G., Rusanov, A., & Weng, C. (2013b). Sick patients have more data: the non-random completeness of electronic health records. Teoksessa *AMIA Annual Symposium proceedings, 2013 (1472–1477).*

Weiskopf, N. G, Bakken, S., Hripcsak, G., & Weng, C. (2017). A Data Quality Assessment Guideline for Electronic Health Record Data Reuse. *EGEMS (Washington, DC)*, *5*(1), 14.

Whittemore, R., Chase, S. K. & Mandle, C. L. (2015). Pearls, Pith, and Provocation. Validity in Qualitative Research. *Qualitative Health Research, 11* (4), 522-537.

Xiong, W., Yu, Z., Bei, Z., Zhao, J., Zhang, F., Zou, Y., … Xu, C. (2013). A Characterization of Big Data Benchmarks, (1), 118–125.

Yin, R. K. (2018). *Case Study Research and Applications. Design and Methods.* (6. painos). Los Angeles: SAGE.

Yoo, Y. (2015). It is not about size: A further thought on big data. *Journal of Information Technology*, *30*(1), 63–65.

Zhu, Y., & Cai, L. (2015). The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal*, *14*(2), 1–10.

LIITE 1 LIITETAULUKKO 1

LIITETAULUKKO 1 Terveydenhuollon tiedon laatuongelmat, niiden syyt ja seuraukset toissijaisessa käytössä kirjallisuuskatsauksen pohjalta

Tiedon laatuominaisuus Wangin ja Strongin (1996) mukaan*	**Laatuongelmien syyt**	**Laatuongelmien seuraukset**
Sisäinen laatu		
Virheettömyys	**Kirjaaminen:** kliinisen hoidon prosessi[14], manuaalisen kirjaamisen virheet[7, 9, 10, 16, 19, 22], työtavat [19, 22, 3], kiire[16, 25], motivaatio[16, 20, 25, 31], kirjaaminen ei ole prioriteetti[9], kirjaamisperuste[9], mittausvirheet[6, 10,] koodaamisen vaativuus[25], työvoiman ja osaamisen puute[25], ohjelmisto- ja ohjelmointivirheet[6,16], yhteen toimimattomat järjestelmät[28], samojen asioiden kirjaaminen moneen paikkaan[28], tietojärjestelmien helppokäyttöisyys[22], automatiikka[22] **Toisiokäyttö:** ei tietoa tiedon alkuperästä ja syntyprosessista [20], laatua pidetään itsestäänselvyytenä[22], tiedon vaihtelevat formaatit ja merkitykset eri tallennus- ja käyttökonteksteissa[19], tiedon yhdenmukaistaminen, erityisesti useat ETL-prosessit[22], entiteettien hajoaminen[22], puuttuvan tiedon käsittely[23], tiedon prosessoinnin ja varastoinnin teknologiat ja menetelmät ja automatisoiminen[22]	**uskottavuuden puute**[6], tiedon laadunvarmistaminen **vaatii paljon manuaalista työtä (teknisen käsittelyn helppous)**[19], systemaattiset harhat ja kausaalipäättelyn ongelmat[10,14], tutkimuksen ja analytiikan tulokset virheellisiä[9,22,29] päätöksenteon virheet[23,] **vertailtavuus**[3]
Objektiivisuus	tieto narratiivisessa muodossa[6], vaikeasti tulkittava tieto[6]	
Uskottavuus	virheellinen tieto[6]	
Saatavuuden laatu		
Saatavuus	järjestelmien yhteentoimimattomuus ja tiedon hajanaisuus[10], tiedon yhdenmukaisuuden puute[26], tietoa ei ole tallennettu rakenteisesti[6], narratiivisen tiedon käytön hankaluudet[9,26], organisaationaaliset, eettiset ja lailliset esteet[8,12,31],	epätäydellisyys **(täydellisyys)**[26]**, vertailtavuus**[3]
Pääsyn tietoturva	tiedon siiloutuneisuus ja tietoturvasäädökset[12]	Tiedon saatavuuden ongelmat **(saatavuus)**[12]
Kontekstuaalinen laatu		
Relevanssi	keruu- ja käyttökontekstien erot[22], metatieto keruukontekstista puuttuu[19, 23], tietoihin tehdyt muunnokset [9], tietoa ole kerätty tutkimusprotokollien mukaan [9], koodauksen puutteet[2], koodistojen puutteet[2], tiedon epätäydellisyys[2, 10, 12, 14]	hyödyntäminen haastavaa tai mahdotonta[9]

(Jatkuu)

Liitetaulukko 1 (jatkuu)

Ajantasaisuus	diagnoositietoja ei kirjata joka käynnillä[9] , potilastietojärjestelmistä ei näe sairauden puhkeamisen ajankohtaa [6]	puuttuva diagnoositieto ei todista sitä, että potilaalla ei ole kyseistä sairautta (**täydellisyys, virheettömyys**)[9]
Täydellisyys	**Kirjaaminen:** kliininen työnkulku[9], tiedon keruu- ja varastointimenetelmät ja rakenne[14], dokumentaatiokäytännöt[9, 21], kirjaaminen ristiriidassa potilaiden hoidon kanssa[25], lääkäri ei pidä tietoja kirjaamisen arvoisina, unohtaa kirjata, ei kysy potilaalta [16], narratiivinen kirjaaminen koodaamisen sijasta[16], kiire[16, 25], lääkärit eivät ajattele kirjatessaan tutkimustarkoituksia[10], lääkäri ei tunne tiedon käyttötarkoituksia[20], tietojärjestelmien helppokäyttöisyys[10], koodaaminen lääkärille vaivalloista verrattuna narratiivisen tekstin kirjoittamiseen[6], potilaat saavat hoitoa eri organisaatioissa tai eivät muuten hakeudu seurantaan[9], potilasta ei kutsuta seurantaan[16], potilaat eivät ole säännöllisesti vuorovaikutuksessa terveydenhuoltoon [6], potilaalla ei pääsyä erikoislääkärille [32], terveet ihmiset käyttävät terveyspalveluita vähemmän kuin sairaat[14], diagnoosia ei ole voitu tehdä sairauden alkuvaiheessa[32] **Toisiokäyttö:** monet tietolähteet, joissa ei yksilöiviä tunnisteita[31], ei tietoalkioiden välisiä yhteyksiä[15], tiedon harmonisaation puute[16], entiteettien hajoaminen[22], tietoa on vaikea hankkia, puhdistaa, aggregoida ja analysoida[26], tiedon poimintamenetelmät[3]	rakenteisen tiedon puuttuessa joudutaan käyttämään rakenteetonta tekstiä[6], joudutaan käyttämään vaihtoehtoisia tiedonlähteitä tai tilastollisia tekniikoita [6], rajoittaa tulosmuuttujia, selittävien tekijöiden määrää ja populaation kokoa **(tiedon määrä)**[6], vasemmalta tai oikealta sensurointi [9, 30], aineisto ei ole kliinisten kokeiden tiedonkeruuprotokollien mukainen [6], yhdistämisen virheet [23], yhdistäminen ei onnistu[31], ei voida käyttää **(relevanssi)**[2, 10, 12, 14], tiedon epäluotettavuus **(virheettömyys)**[10], edustavuuden puute[14], systemaattiset harhat ja kausaalipäättelyn ongelmat[10] **vertailtavuus**[3]
Tiedon määrä	sairauskertomustiedon kenttien epätäydellisyys voi rajoittaa mukaan otettavan populaation kokoa[6]	
Representationaalinen laatu		
Tulkittavuus	tieto kirjattu käyttäen erilaisia, usein laadullisia, määritelmiä[6]	**objektiivisuuden puute**[6], kyseenalainen vertailtavuus **(vertailtavuus)**[6]
Esittämisen tiiviys	rakenteinen tieto puuttuu[6]	
Esittämisen yhdenmukaisuus	standardoinnin puute[15], koodistojen ja standardien erot ja muutokset[22], tietojärjestelmien keräämä tietosisältö vaihtelee [1, 5, 16, 18], eri tietojärjestelmät eivät semanttisesti yhteentoimivia[13, 16, 17, 25, 26, 30, 33], eri palveluntarjoajat keräävät eri yksityiskohtia ja eri aikoina[6], eri tallennus- ja käyttökontekstit[19], koodaamiskäytäntöjen muutokset[9, 22], terminologian puute[2], kirjaamisen ohjeistus[19], kirjaamisperusteet[20], lääkäri ei tiedä tiedon käyttötarkoituksia[20], narratiivinen teksti[17]	vaikeuttaa tulosten tulkintaa, virhetulkinnat [9], vaikeuttaa tiedon käyttöä, laadunvarmistamisen manuaalinen työ **(teknisen käsittelyn helppous)**[19], tietojen yhdistäminen vaikeaa, entiteettien hajoaminen [22,31], tieto epäluotettavaa **(virheettömyys)**[10], potilaan tiedot hajallaan[16], mikä hidastaa

(Jatkuu)

Liitetaulukko 1 (jatkuu)

		tiedon kulkua ja työntekoa sekä tarvetta kirjata samoja tietoja moneen paikkaan ja altistaa virheille **(virheettömyys)** [28], tietoa on vaikea jakaa organisaatioiden välillä, ja tietoa on vaikea hankkia **(saatavuus),** puhdistaa, aggregoida ja analysoida **(teknisen käsittelyn helppous)** [26], tietojen epätäydellisyys **(täydellisyys)**[6, 24], vaikuttaa alentavasti sekä tiedon **virheettömyyteen**[19] että **täydellisyyteen**[16], vaikeaa aggregoida ja **analysoida (teknisen käsittelyn helppous)**[23], tieto käyttökelvotonta **(relevanssi)**[15], voi johtaa virheellisyyteen **(virheettömyys)** tai tutkimuksen harhaan, jos sitä ei oteta huomioon[6,] ei voida tehdä kohortti- tai väestötason analyysejä täydellisellä aineistolla **(täydellisyys)** [24]
Mallista pois jätettyjä laatuominaisuuksia		
Jäljitettävyys	käyttäjä ei tunne tiedon lähteitä[20], saman tiedon monet lähteet, esim. lääkitys [9], tieto sairauskertomuskertomusjärjestelmän rakenteen ja ominaisuuksien vaikutuksista tietoon puuttuu[11], metatiedon puute[19, 23], tietomallit eivät ole läpinäkyviä ja täsmällisiä kaupallisista syistä[4]	kausaalipäättelyn ongelmat[11]**, relevanssi**[19, 23], **virheettömyys**[4]
Kustannustehokkuus	tiedon louhiminen narratiivisesta tekstistä vie aikaa ja resursseja[6]	louhiminen ei aina kannattavaa **(saatavuus)**[6]
Teknisen käsittelyn helppous	semanttisen yhdenmukaisuuden puute[19], rakenteetonta tietoa on vaikea puhdistaa, aggregoida ja analysoida [26], manuaalisen prosessoinnin tarve laadunvarmistuksessa[19], viiteavainten ja yksilöivien tietojen puute [22, 31]	tietoa ei pystytä käyttämään **(saatavuus)**[9], tietoa vaikea jakaa organisaatioiden välillä ja hankkia **(saatavuus)**[26], **kustannustehokkuus**[6]
Muita laatuominaisuuksia		
Tarkkuus	laskutustarkoituksiin kerätyt diagnoosit hyvin epätarkkoja [9], ICD-9 -diagnoosikoodisto ei erottele syöpäpotilaiden alkuperäisiä ja metastaattisia kasvaimia[2], rakenteeton tieto[17]	**epätäydellisyys**[2]
Yhtenäisyys	yhteen toimimattomat tietojärjestelmät tekevät tiedosta hajanaista[10]	Tiedon saatavuuden ongelmat **(saatavuus)**[10]
Rakenteisuus	tietojärjestelmien käytettävyys, tieto kerätty narratiivisena tekstinä[6]	rajallinen hyöty tutkimukselle **(relevanssi)**[6] epätäydellisyys **(täydellisyys)**[6] riittämätön tarkkuustaso **(tarkkuus)**[17]

(Jatkuu)

Liitetaulukko 1 (jatkuu)

		objektiivisuuden puute **(objektiivisuus)**[6] tiiviys puuttuu **(esittämistavan tiiviys)**[6], epäyhdenmukaisuus **(yhdenmukaisuus)** [6, 17], teknisen käsittelyn hankaluus **(teknisen käsittelyn helppous)**[26], kustannustehokkuuden puute **(kustannustehokkuus)**[6]
Vertailtavuus	tietoja ei ole tallennettu yhdenmukaisella tavalla[6, 19], tiedon huono tulkittavuus[6], tietojen tallennustavoista eri aikoina eri paikoissa ei ole tietoa[6, 19], sähköisen sairauskertomusjärjestelmän sisältö, rakenne ja formaatti, paikalliset tiedonkeruun ja -poiminnan menettelytavat, järjestelmissä käytettyjen koodistojen ja menuvalintojen vaihteleva tarkkuustaso, erilaiset tiedon laatustandardit, tietolähteiden valinta [3]	erojen tulkitseminen muutokseksi mahdotonta **(relevanssi)**[19], virheellisten tulkintojen riski[3], kielteiset vaikutukset hoidon laadun vertailevalle ja hoidon tuloksellisuuden tutkimukselle **(lisäarvo)**[3], estää tiedon toissijaista käyttöä **(lisäarvo)**[27]

*Taulukosta on jätetty pois maine, lisäarvo, ymmärtämisen helppous ja tiedonlähteiden moninaisuus, sillä niihin ei ollut selviä viittauksia. Taulukkoon on lisätty Wangin ja Strongin (1996) viitekehykseen tai siitä pois jätettyihin muuttujiin kuulumattomat laatuominaisuudet tarkkuus, yhtenäisyys ja rakenteisuus.

Lähteet: Kane, 1997[1]; Botsis ym., 2010[2]; Chan ym., 2010[3]; Liaw ym., 2011[4]; Estabrooks, 2012[5]; Bayley ym., 2013[6]; Borycki, 2013[7]; Buhl ym., 2013[8]; Hersh ym., 2013[9]; Hoffman & Podgurski, 2013[10]; Holve ym., 2013[11]; Jee & Kim, 2013[12]; Liaw ym., 2013[13]; Weiskopf ym., 2013b[14]; Dentler ym., 2014[15]; Hoffman, 2014[16]; Raghupathi & Raghupathi, 2014[17]; Richesson ym., 2014[18]; Laine ym., 2015[19]; Markus & Topi, 2015[20]; Nobles ym., 2015[21]; Sukumar ym. 2015[22]; Clarke, 2016[23]; Dinov, 2016[24]; Dungey ym., 2016[25]; Kruse ym., 2016[26]; Sáez ym., 2016[27]; Hyppönen ym., 2018[28]; Pitoglou, 2018[29]; Jetley & Zhang, 2019[30]; Molinari & Nollo, 2020[31]; Piri, 2020[32]; Rayner ym., 2020[33]

LIITE 2 HAASTATTELURUNKO

1. **Aloitus**
    - Haastateltava, aika, paikka
    - Haastattelun aihe: Tiedon hyödyntäminen tiedon laadun näkökulmasta ”Urologian hoitopolku -hankkeessa”
    - Varmistaminen, että haastateltava on saanut tiedotteen ja tietosuojaselosteen ja on tietoinen osallistumisen vapaaehtoisuudesta

2. **Taustatiedot**
    - Ammatti ja työtehtävät
    - Koulutus ja kokemus
    - Rooli Urologian hoitopolku-hankkeessa?
    - Käyttökokemus aineistosta?
    - Mitkä ovat tämän hankkeen tavoitteet ja mitä siinä tehdään?

    Jos ei mukana urologian hankkeessa:
    - Rooli VSSHP:n tietoallasasioissa?
    - Tunnetko Urologian hanketta?
    - Miten olet tekemisissä potilasrekisteritiedon kanssa?

3. **Kliinikko: Mihin tarkoituksiin käytät Urologian hoitopolku -aineistoa?** (tai käsitys sen käytöstä)
    - Mitä tarkoittaa tai sisältää käytännössä?
    - Millaisia asioita selvität aineiston avulla?
    - Mitä tiedolla tehdään?
    - Mikä tiedon käyttämisen tavoitteena?
    - Mitä hyötyä on tiedon käytöstä?
    - Tärkeimmät käyttötarkoitukset?
    - Hoidon, tiedolla johtamisen ja tutkimuksen **suhde toisiinsa**, mitä samaa, **ristiriitaa**, prioriteettijärjestys?

4. **Mitä tietoja tähän tarkoitukseen tarvitaan?**

5. **Mitä vaatimuksia näille tiedoille on, että voit käyttää niitä tähän tarkoitukseen? TAI** Mitä **vaatimuksia on tietoaltaan potilasrekisteritiedoille**, että niitä voidaan käyttää urologian hankkeen tyyppisessä tarkoituksessa (hoidon laadun kehittäminen ja seuranta)?

6. Puhutaan edelleen näistä **urologian klinikan potilaiden tiedoista ja niiden käyttämisestä** (päätöksentekoon työssäsi). **Piirtäisitkö kuvan**, joka kuvaa prosessia tiedon tallentamisesta siihen, kun käytät sitä (tai sitä käytetään) päätöksentekoon (työssäsi). Voit keskittyä erityisesti niihin vaiheisiin, jotka tunnet oman työsi kautta parhaiten….Kerro, mitä olet piirtänyt.

7. **Tiedon hyödyntämisen ongelmat ja niiden syyt**
   Mitä ongelmia on Urologian hoitopolku -tiedon hyödyntämisessä tähän tarkoitukseen (tärkein tarkoitus, joka edellä mainittu)?

   - Suurin ongelma?
   - Mitä tarkoittaa käytännössä (esimerkki/tarkemmalla tasolla)?
   - Miksi se on ongelma?
   - Mitä siitä seuraa?
   - Mistä vaiheesta prosessia (ks. piirretty kuva) ongelma on lähtöisin?
   - Mistä syistä tämä johtuu?

8. **Haluatko vielä tuoda esiin jotain aiheeseen liittyvää?**