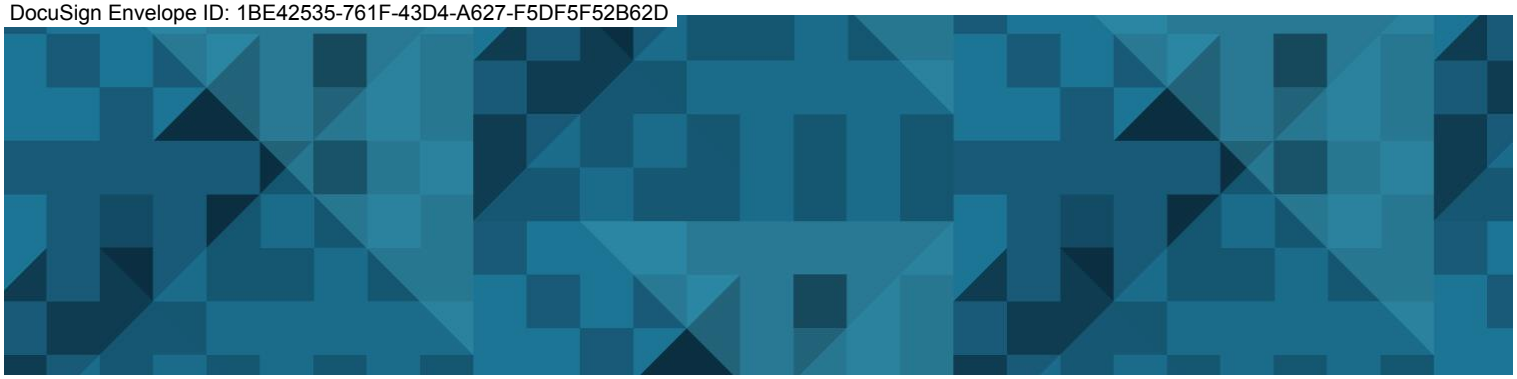
VTT Technical Research Centre of Finland

Toisiolain vaikutukset tutkimukseen ja data-analytiikan sovelluksiin

Pajula, Juha; Viiri, Sampo; Similä, Heidi; Lähteenmäki, Jaakko; Tuomi-Nikula, Antti

Published: 08/02/2021

Document Version
Publisher's final version



TUTKIMUSRAPORTTI

VTT-R-00118-21

Toisiolain vaikutukset tutkimukseen ja data-analytiikan sovelluksiin

Hyteairon analytiikkatyöryhmän selvitys

Kirjoittajat: Juha Pajula, VTT
Sampo Viiri, THL
Heidi Similä, VTT
Jaakko Lähteenmäki, VTT
Antti Tuomi-Nikula, THL

Luottamuksellisuus: julkinen

Raportin nimi

Toisiolain vaikutukset tutkimukseen ja data-analytiikan sovelluksiin - Hyteairon analytiikkatyöryhmän selvitys

Asiakkaan nimi, yhteyshenkilö ja yhteystiedot	**Asiakkaan viite**
Sosiaali- ja terveysministeriö, Jukka Lähesmaa, jukka.lahesmaa@stm.fi Hyteairo-analytiikkaverkosto, Maritta Perälä-Heape, maritta.perala-heape@oulu.fi	VN/23904/2020
Projektin nimi	**Projektin numero/lyhytnimi**
Toisiodataselvitys	ToDa
Raportin laatija(t)	**Sivujen/liitesivujen lukumäärä**
Juha Pajula, Sampo Viiri, Jaakko Lähteenmäki, Heidi Similä, Antti Tuomi-Nikula	23/7
Avainsanat	**Raportin numero**
Terveysdata, toisiokäyttö, Findata, etäkäyttöympäristö	VTT-R-00118-21

Tiivistelmä

Laki sosiaali- ja terveystietojen (sote-tiedot) toissijaisesta käytöstä (552/2019, ns. toisiolaki) tuli voimaan keväällä 2019 ja sen säännösten soveltaminen laajenee vaiheittain. Osana Hyteairo-ohjelmaa sosiaali- ja terveysministeriö (STM) tilasi tämän raportin tuottaneen selvitysprojektin, jonka tavoitteena oli kartoittaa toisiolain ja siihen liittyvien määräysten vaikutusta tekoälyteknologioihin liittyvään tutkimustoimintaan. Kartoitus perustui haastatteluihin, joihin osallistui alan toimijoita aina valtion laitoksista yliopistoihin ja dataa hyödyntäviin yrityksiin saakka. Projektissa tarkasteltiin myös vaihtoehtoja uuden lain mukaisten tietoturvallisten laskentaympäristöjen kehittämiseen. Erityisesti selvitettiin, millaisia haasteita ja mahdollisuuksia liittyy laskentaympäristöjen toteutukseen toisiolain uuden määräyksen (5.10.2020) puitteissa. Työryhmä koostui VTT:n ja THL:n asiantuntijoista. Haastattelujen lisäksi hyödynnettiin saatavilla olevaa julkista materiaalia sekä projektityöryhmän aiempaa kokemusta tekoälyn potentiaalista tutkimuksessa ja ratkaisuista sote-tiedolla johtamisessa.

Projektin tavoitteena oli selvittää toisiolain tuomien uusien vaatimusten ja toimintamallien vaikutusta käytäntöön, erityisesti, kun kyseessä on tekoälytutkimus. Raportti keskittyy potentiaalisten teknologioiden ja ratkaisujen tunnistamiseen ja niihin liittyvien rajoitteiden selventämiseen. Lisäksi raportissa haetaan näkökulmia uusien ratkaisujen tukemiseen, niin lupien kuin myös tietoturvallisten käyttöympäristöjen kehittämisen kannalta. Selvityksen osana esitetään jatkotoimenpidesuosituksia tutkimuksen tukemiseen, datan luvittamiseen sekä käyttöympäristöjen määrittelyyn ja toteutukseen tehokkaan ja sujuvan tutkimuskäytön varmistamiseksi. Raportti toimii pohjana päätettäessä tarvittavista toimenpiteistä datan toisiokäytön potentiaalin hyödyntämisessä nyt ja tulevaisuudessa sekä lähtökohtana STM:n tuleville selvityksille sote-tiedon hyödyntämisestä toisiolain puitteissa tiedolla johtamisen ja resursoinnin näkökulmasta.

Luottamuksellisuus	julkinen

Tampere 8.2.2021

Laatija	**Tarkastaja**
Juha Pajula, Research Scientist	Jari Ahola, Research Team Leader

VTT:n yhteystiedot

Juha Pajula, +35840 163 0719, juha.pajula@vtt.fi, P.O. Box 1300, 33101, Tampere, Finland

Jakelu (asiakkaat ja VTT)

STM, VTT, THL ja muu jakelu.

Hyväksyminen

Päivämäärä:	09 helmikuuta 2021
Allekirjoitus:	DocuSigned by: Kari Kohtamäki A8634D9764764A2...
Nimi:	Kari Kohtamäki
Asema:	Customer Account Lead

Alkusanat

Sosiaali ja terveystietojen toisiokäyttö on digitaalisen transformaation tukipilari, minkä edelläkävijänä Suomi tunnetaan. Suomessa on luotu valmiuksia sote-tiedon saatavuuden ja lupamenettelyjen yhtenäistämiselle, mikä luo hyvän perustan datan toisiokäyttöön tutkimuksessa ja innovaatiotoiminnassa. Datan saatavuuden lisäksi on luotava valmiuksia sen hyödyntämiselle, erityisesti uusien palvelujen ja tuotteiden kehityksessä. Edistyksellinen data-analytiikka ja tekoäly/koneoppimisen menetelmät tuovat merkittävää uutta näkökulmaan tulevaisuuden tiedolla johtamisen malleihin. Tekoälyllä ja robotiikalla odotetaan olevan merkittäviä hyötyjä hyvinvointialan palveluissa ja niiden tuottamisessa.

Siksi Hyteairo-ohjelmassa ”Tekoäly analytiikassa” on nostettu yhdeksi yhteistyötä mahdollistavaksi kehittämiskokonaisuudeksi. Hyteairo-ohjelma on kaikkien osapuolien yhteinen ohjelma yhteydenpitoon ja hyvinvointialan uudistamiseen tekoälyn ja robotiikan avulla.

Hyteairo-ohjelman eri verkostot selvittävät ja kehittävät näitä mahdollisuuksia. Tekoäly-analytiikassa verkoston tavoitteena on saattaa kehittämistyön parissa työskenteleviä organisaatiota ja ihmisiä yhteen ja luoda yhteinen käsitys datan tehokkaammasta käytöstä, tekoäly-avusteisesta tietojohtamisesta sekä data-pohjaisten palvelumallien kehityksestä. Verkosto tekee selvityksiä, tuottaa näkemyksen tekoälyn potentiaalista ja rakentaa aihealueen kansallista verkostomaista yhteistyötä.

Tämä raportti sisältää kuvauksen ja tulokset sosiaali- ja terveysministeriön Hyteairo-analytiikka työryhmältä tilaamasta selvityksestä toisiolain vaikutuksista tutkimukseen ja data-analytiikan sovelluksiin. Selvityksen toteutti VTT:n ja THL:n asiantuntijoista koottu projektiryhmä haastattelujen ja kirjallisuusselvityksen avulla.

Tämän selvityksen tavoitteena oli haastattelututkimuksen ja kirjallisuusselvityksen kautta kartoittaa Suomen sosiaali-ja terveysdataan kohdistuvan tekoälytutkimuksen nykytilannetta ja siihen liittyviä rajoituksia. Selvityksessä tarkasteltiin toisiolain vaikutuksia julkisen ja kaupallisen tekoälyä hyödyntävän tutkimuksen toteuttamiseen ja toisiolain mukaisten tietoturvallisten käyttöympäristöjen kehittämiseen huomioiden tekoälyä hyödyntävien tutkimusprojektien käytännön tarpeet. Tämä raportti keskittyy tekoälyn hyödyntämiseen tutkimuksessa ja kehittämisessä.

Se toimii samalla lähtökohtana tulevia tiedolla ohjaukseen ja johtamiseen liittyviä selvityksiä silmällä pitäen. Raportin havaintojen perusteella projektitiimi antaa myös muutamia suosituksia suurimpien haasteiden poistamiseksi ja toisiolain alaisen toiminnan täyden potentiaalin saavuttamiseksi.

Projektin johtoryhmään kuuluivat Jukka Lähesmaa (STM), Maritta Perälä-Heape (Hyteairo / Oulun Yliopisto), Antti Piirainen (Findata) ja Juuso Rahkola (KELA).

Projektiryhmä kiittää kaikkia haastateltuja ja ohjausryhmää onnistuneesta selvityksestä.

Tampereella ja Helsingissä 8.2.2021,

Juha Pajula,
Sampo Viiri,
Jaakko Lähteenmäki,
Heidi Similä ja
Antti Tuomi-Nikula

Sisällysluettelo

1. Johdanto

Laki sosiaali- ja terveystietojen (sote-tiedot) toissijaisesta käytöstä[1] (552/2019, ns. toisiolaki) on tullut voimaan keväällä 2019 ja sen säännösten soveltaminen laajenee vaiheittain. Lainsäädäntö vaikuttaa korkeakouluissa, tutkimuslaitoksissa ja yrityksissä toteutettavaan sosiaali- ja terveysalan tietoaineistoja hyödyntävään tutkimustoimintaan ja opetukseen. Alan tutkijat ovat eri yhteyksissä tuoneet esiin näkemyksensä, että vaikka lain tavoite on hyvä, sen säännökset monella tavalla monimutkaistavat alan tutkimuksen käytänteitä. Toisiolakia on sovellettu keväästä 2020 alkaen, ja sen tulkintaan liittyy vielä epäselvyyksiä, erityisesti tilanteissa joissa datan käyttöön liittyy myös muita lakeja kuten tilasto- ja biopankkilait.

Hyteairo-ohjelman analytiikka-alatyöryhmä[2] on perustettu syksyllä 2019. Alatyöryhmä koostuu eri organisaatioiden nimeämistä jäsenistä (Kela, THL, Findata, VTT, DigiFinland Oy, Metropolia, KSSHP, FCAI) ja sen koordinaatiovastuu on Oulun yliopistolla (Lääketieteellinen tiedekunta). Työryhmä edistää kansallisten tietojohtamisen strategioiden toimeenpanoa sekä tutkimus-, kehittämis- ja innovaatiotoiminnan vahvistamista. Sen tehtävänä on selvittää tekoälyn hyödyntämismahdollisuudet osana toisiolain toimeenpanoa. Lisäksi sen tehtävänä on näkemyksen tuottaminen tekoälyn potentiaalista sekä osaamistarpeista analytiikassa. Kansallinen data-analytiikkaosaamisen kehittäminen ja yhdessä oppiminen asettuvat luontevasti osaksi käynnissä olevia ohjelmia ja kehittämishankkeita, kuten esimerkiksi toisiolain toimeenpanoa ja sote-tiedolla johtamisen Valtava-ohjelmaa ja Toivo-hanketta[3].

Osana Hyteairo-ohjelmaa sosiaali- ja terveysministeriö (STM) tilasi tämän raportin tuottaneen selvitysprojektin, jonka tavoitteena oli kartoittaa toisiolain ja siihen liittyvien määräyksien vaikutusta tekoälyteknologioihin liittyvään tutkimustoimintaan. Kartoitus perustui haastatteluihin, joihin osallistui alan toimijoita aina valtion laitoksista yliopistoihin ja dataa hyödyntäviin yrityksiin saakka. Lisäksi projektin tavoitteena oli tarkastella vaihtoehtoja uuden lain mukaisten tietoturvallisten laskentaympäristöjen kehittämiseen. Erityisesti selvitettiin millaisia haasteita ja mahdollisuuksia liittyy laskentaympäristöjen toteutukseen toisiolain uuden määräyksen[4] (5.10.2020) puitteissa kolmansien osapuolien toimesta.

Työryhmä koostui VTT:n ja THL:n asiantuntijoista. Haastattelujen lisäksi työssä hyödynnettiin saatavilla olevaa julkista materiaalia sekä projektityöryhmän aiempia selvityksiä tekoälyn mahdollisuuksista tutkimuksessa ja ratkaisuista sote-tiedolla johtamisessa.

Sosiaali- ja terveysalan tietolupaviranomainen Findata on yksi keskeinen toisiolain toimija. Se myöntää luvat sosiaali- ja terveystietojen toissijaiseen käyttöön silloin, kun tietoja yhdistellään usealta rekisterinpitäjältä, kun rekisteritiedot ovat peräisin yksityisiltä sosiaali- ja terveydenhuollon palvelunjärjestäjiltä tai kun kyse on Kanta-palveluihin tallennetuista tiedoista (vuodesta 2021 eteenpäin). Toisiolain 6 §:ssä on määritelty palveluista vastaavat viranomaiset ja organisaatiot sekä tietoaineistorajaukset. Findata ei siis myönnä lupia kaikkiin rekisterinpitäjien aineistoihin.

Lokakuussa 2020 toisiolakiin annettiin erillinen Findatan määräys tietoturvallisten laskentaympäristöjen toteutuksesta ja ylläpidosta. Määräykseen liittyvissä lausunnoissa esitettiin huolta määräyksen vaikutuksista tutkimustoiminnan käytännön mahdollisuuksiin. Koska määräys on uusi, projektissa pyrittiin haastattelujen kautta selvittämään miten eri toimijat näkevät määräyksen vaatimukset ja miten siihen aiotaan reagoida. Samalla kävimme

[1] https://www.finlex.fi/fi/laki/alkup/2019/20190552
[2] https://thl.fi/fi/tutkimus-ja-kehittaminen/tutkimukset-ja-hankkeet/hyvinvoinnin-tekoaly-ja-robotiikka-ohjelma-hyteairo-/analytiikkaosaamisverkosto
[3] https://soteuudistus.fi/toivo-ohjelma
[4] https://www.findata.fi/uploads/2020/10/20ddc0dd-findata-maarays-1-2020-muiden-palveluntarjoajien-tietoturvallisille-kayttoymparistoille-asetettavat-vaatimukset.pdf

keskustelua siitä, millaisia resursseja tällaisessa ympäristössä pitäisi olla saatavilla, jotta tekoälyratkaisuja voidaan kehittää ja tutkia ilman merkittäviä rajoitteita.

Findatan lisäksi keskeisiä toimijoita ovat erityisesti data-aineistojen rekisterinpitäjät, joiden toimintaan toisiolaki oleellisesti vaikuttaa. Rekisterinpitäjät antavat toisiolain mukaan neuvontaa omia aineistojaan koskien. Lisäksi toisiolaki velvoittaa rekisterinpitäjät poimimaan aineistot tietyssä ajassa ja vastaamaan Findatan aineistokyselyihin ja hinta-arviopyyntöihin.

THL:n projektiin osallistuneella työryhmällä on pitkä kokemus tutkimusdatan luvittamisesta ja valmistelusta alan tutkimuksiin. Projektiryhmään kuuluvat asiantuntijat toimivat myös nykyään yhdyskanavana Findatan suuntaan hankkeissa, joissa Findata luvittaa THL:n aineistoja toisiolain valtuutuksella tutkimusosapuolille. Yleisesti THL:n rekisteriaineistot ovat keskeisimpiä toisiolain piiriin kuuluvia aineistoja ja ne ovat jo pitkään olleet tärkeä lähde alan tutkimukselle. THL:n oma tutkimus- ja viranomaistoiminta ei sinällään ole riippuvainen Findatasta, mutta THL toimittaa dataa toisiolain mukaisiin hankkeisiin Findatan tietolupien perusteella.

VTT:n keskeinen rooli oli tuoda projektiin tutkimuslaitosten ja kaupallisten toimijoiden näkökulma datan hyödyntämiseen erityisesti toisiolain soveltamisen ja tietoturvallisten ympäristöjen näkökulmasta. VTT:n tutkimusryhmän jäsenet ovat olleet mukana useissa projekteissa, joissa hyödynnetyt data-aineistot ovat Findatan lupien alaisia. VTT on tehnyt ja valmistellut viimeisten vuosien aikana useita hankkeita, jotka ovat toisiolain piirissä tai siihen liittyvien toimintojen parissa. Näistä esimerkkinä on PreMed[5] (”Data-driven precision medicine ecosystem”), jossa yhdisteltiin dataa kolmesta suomalaisesta biopankista sekä kansallisista rekistereistä. Projekti käynnistettiin ennen toisiolain voimaantuloa ja toimii näin hyvänä verrokkina toisiolain vaikutuksia arvioitaessa. Osa projektiryhmän jäsenistä on mukana myös PreMed-hankkeessa. PreMed päättyy keväällä 2021.

Tässä projektissa oli tarkoituksena selvittää uusien vaatimusten ja toimintamallien vaikutusta käytäntöön, erityisesti, kun kyseessä on tekoälytutkimus, jossa tarvittavat aineistot ovat yleensä merkittävästi suurempia kuin perinteisissä rekisteritutkimuksissa. Tämä raportti keskittyy potentiaalisten teknologioiden ja ratkaisujen tunnistamiseen ja niihin liittyvien rajoitteiden selventämiseen. Lisäksi raportissa haetaan näkökulmia uusien ratkaisujen tukemiseen, niin lupien kuin myös tietoturvallisten käyttöympäristöjen kehittämisen kannalta. Selvityksen osana esitetään jatkotoimenpidesuosituksia tutkimuksen tukemiseen, datan luvittamiseen sekä käyttöympäristöjen määrittelyyn ja toteutukseen tehokkaan ja sujuvan tutkimuskäytön varmistamiseksi. Jatkotoimenpidesuositukset kohdistettiin selvityksessä tehtyjen havaintojen perusteella. Raportti toimii näin mahdollisena pohjana päätettäessä tarvittavista toimenpiteistä datan toisiokäytön potentiaalin hyödyntämisen nyt ja tulevaisuudessa.

[5] https://www.vtt.fi/premed

2. Tavoite

Tämän selvityksen tavoitteena oli haastattelututkimuksen ja kirjallisuusselvityksen kautta kartoittaa Suomen terveysdataan kohdistuvan tekoälytutkimuksen nykytilannetta ja siihen liittyviä rajoituksia. Selvityksessä tarkasteltiin toisiolain vaikutuksia julkisen ja kaupallisen tekoälyä hyödyntävän tutkimuksen toteuttamiseen ja toisiolain mukaisten tietoturvallisten käyttöympäristöjen kehittämiseen huomioiden tekoälyä hyödyntävien tutkimusprojektien käytännön tarpeet. Selvitykselle oli konkreettinen tarve, sillä esimerkiksi Hyteairo-analytiikkatyöryhmän keskuudessa on noussut esiin epävarmuutta liittyen toisiolain tulkintaan ja toteutukseen sekä tekoälytekniikoiden käyttöön pelkästään Findatan käyttöympäristöissä. Samalla selvityksellä kartoitettiin alalla tällä hetkellä potentiaalisiksi nähtyjä aiheita ja tutkimuskohteita, joita nykyiset datan luvituksen ja käytön reunaehdot mahdollisesti saattaisivat rajoittaa.

Yleisesti on tunnustettu, että keskeisimpiä lähtökohtia tekoälyä hyödyntävien projektien osalta ovat datan saavutettavuus sekä mahdollisuus riittävän laskentakapasiteetin käyttöön datan käsittelyssä[6]. Modernit pilviratkaisut mahdollistavat suuren laskentakapasiteetin, mutta niiden käytöstä sensitiivisen datan kanssa on esitetty julkisuudessa myös epäilyksiä[7]. Findatan määräys ulkopuolisten toimijoiden ylläpitämistä tietoturvallisista käyttöympäristöistä tuo tilanteeseen uusia mahdollisuuksia. Projektissa kartoitettiin, millaisia vaihtoehtoja teko-älytutkimukselle todellisuudessa on toisiolain piirissä ja mitkä ovat keskeisimmät rajoitteet nykytilanteessa.

Haastattelujen ja selvitysten perusteella selvityksessä tuotettiin näkemys siitä, miten tekoälyä ja data-analytiikkaa voidaan hyödyntää ja kehittää suomalaisella terveysdatalla ja mitkä ovat mahdollisesti haastattelujen perusteella potentiaalisia aiheita lähitulevaisuudessa. Selvitys keskittyy tutkimus- ja innovaatiotoimintaan huomioiden erilaiset tietoturvavaatimukset ja käyttöympäristöt toisiolain piirissä.

[6] https://doi.org/10.1038/s42256-020-0186-1
[7] https://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/162451/VM_2020_66.pdf

3. Toteutus

Selvitys toteutettiin ensisijaisesti haastattelututkimuksena, jonka havaintoja peilattiin projektitiimin omiin kokemuksiin toisiolain piirissä tapahtuneiden valmisteluihin parissa sekä vanhempiin projekteihin, jotka olisivat nykyään toisiolain piirissä. Tällaisia ovat esimerkiksi PreMed[8] -projekti, joka toimii hyvänä esimerkkinä tilanteesta ennen toisiolain voimaantuloa. Toisiolain alaisista tutkimuksista esimerkkejä ovat esimerkiksi Business Finlandin rahoittama Stroke-Data[9] -projekti ja Okva[10] -projekti. Stroke-Data -tutkimuksessa VTT on keskeisessä roolissa dataan liittyvien lupien vaatimien dokumenttien valmistelussa. Okva-projektissa (Espoon osaamiskeskuksen vaikutusten arviointi) VTT:n projektiryhmä on joutunut selvittämään Findatan ja Tilastokeskuksen välistä työnjakoa tiedon toisiokäytön ja tilastotutkimuksen välillä. Tämän lisäksi selvityksessä tarkasteltiin, millaisia haasteita liittyy Findatan tietoturvallisen tutkimusympäristön käyttöön ja kolmansien osapuolien laskentaympäristöjen toteutusten realisoitumiseen eri datan käyttäjien parissa.

Selvitystyöryhmä haastatteli alan toimijoita, joihin kontaktipinta löytyi projektitiimin omien verkostojen sekä Hyteairon analytiikkaverkoston kautta. Haastateltavat tahot pyrittiin valitsemaan kattavasti datan toisiokäytön toimintaympäristön eri puolilta. Haastatteluilla pyrittiin kattamaan erityisesti julkisen sektorin, tutkimuslaitosten, yliopistojen ja yrityskentän näkökulmat. Tässä selvityksessä tarkastellaan näin ollen toisiolain luomaa terveysdatan käytön tilannetta näiden toimijoiden näkökulmasta. Selvitys ei ota kantaa sote-palveluntuottajien tietojohtamisen kehittämiseen, jossa hyödynnetään vain omia aineistoja, ja joihin ei Findatan lupaa tarvita. Selvitys ei myöskään ota kantaa tietoturvallisten etäkäyttö-ympäristöjen teknisiin ratkaisuihin, vaan tarkastelee näiden toteutusta ja käyttöä käytännön ja tulevaisuuden tarpeiden kannalta.

3.1 Haastattelut

Kaiken kaikkiaan projektitiimi teki 20 haastattelua, joihin osallistui 16 eri organisaation edustajia. Tämän lisäksi projektitiimi kävi erillistä keskustelua Findatan edustajien kanssa Findatan roolista toisiolain toteuttajana. Haastatellut tahot sisälsivät kaksi verkostoa (Hyteairo, FCAI), kolme yliopistoa (Oulun yliopisto, Aalto yliopisto ja Helsingin yliopisto), kaksi julkishallinnon toimijaa (Kela, THL), tutkimuslaitoksen (VTT), kaksi sairaanhoitopiiriä (VSSHP, HUS) sekä kuusi data-analytiikkaa tekevää yritystä ja tietotekniikka-palvelutoimittajan (CSC).

Haastatteluja varten määriteltiin kysymyslista koko selvityksen aihepiirin laajuudelta. Tavoitteena ei kuitenkaan ollut saada kaikilta haastateltavilta vastausta kaikkiin kysymyksiin, vaan ainoastaan heidän kannaltaan relevantteihin kohtiin. Haastattelukysymykset yhteenvetoineen löytyvät liitteistä 1 ja 2. Käytännössä Covid-19 -tilanteen takia haastattelut järjestettiin Teams-kokouksina ja kysymyslista toimitettiin ennakkoon haastateltaville. Haastattelijoita oli Teams-kokouksissa paikalla yksi tai kaksi, joista jälkimmäisessä asetelmassa toinen haastattelijoista toimi yleensä kirjurina. Haastattelut myös tallennettiin väliaikaisesti haastateltavien luvalla yhteenvetojen tarkastelua varten, erityisesti, jos haastattelijoita oli vain yksi. Tallenteet hävitettiin projektin päätteeksi, kun yhteenvedot haastatteluista olivat valmiit. Haastatteluissa kysyttiin myös erikseen, saako haastateltavan nimeä ja edustamaa tahoa mainita projektin julkisissa esityksissä. Kaksi haastattelua tehtiin kirjallisesti siten, että haastateltava vastasi kysymyksiin sähköpostitse.

[8] https://projectsites.vtt.fi/sites/premed/
[9] https://www.mai.fi/ajankohtaista/strokedata-konsortiolle-lahes-kuusi-miljoonaa-tutkimusrahoitusta.html
[10] https://www.espooskillscentrestudy.fi/

3.2 Tekninen selvitys

Toisiolain vaatimukset ja terveysdatan liittyvät muut määräykset luovat uudenlaisia haasteita ja toisaalta mahdollisuuksia datan toisiokäytön ympärille. Keskeinen muutos on se, ettei perinteinen toimintatapa, jossa datan omistaja luovuttaa datan suoraan fyysisesti tutkijalle itselleen ole enää mahdollinen, vaan data luovutetaan Findatan vaatimusten mukaiseen tietoturvalliseen käyttöympäristöön. Muutos vaikeuttaa perinteisesti tehtyä datan monipuolista hyödyntämistä ja aiheuttaa vastarintaa uuden toimintatavan käyttöönottoon. Viimeaikaisten tietovuotojen valossa, kuten psykoterapiakeskus Vastaamon tapauksessa[11], on kuitenkin selvää, ettei terveysdataa voida siirtää ja käsitellä ilman yhtenäistä, jäljitettävää ja turvallista toimintatapaa.

Edellisestä johtuen jatkossa korostuu laskentaympäristöjen toimittajan rooli. Toisiolain mukaisesti dataa tulee säilyttää keskitetysti ja kaikki datan käyttö pitää kirjata jäljitettävyyden ja tietosuojan nimissä, jolloin on selvää, ettei perinteinen henkilökohtaisella tietokoneella tehtävä analytiikka tule kyseeseen kuin poikkeustapauksissa. Näin ollen sensitiivistä tietoa – kuten henkilökohtaisia terveystietoja – ei tule analysoida laitteilla, joita voi helposti siirtää ja joihin pääsee helposti fyysisesti käsiksi (kannettavat ja pöytäkoneet toimisto- ja matkustusympäristöissä). Käytännössä nämä vaatimukset johtavat keskitettyyn datan tallennukseen ja hallintaan sekä tarpeeseen saada datan rinnalle tarpeen mukaan pystytettävää laskentakapasiteettia. Nykyteknologian näkökulmasta tällaiset ratkaisut ovat yleensä pilvi- tai hybridipilviratkaisuja, eli kaikki tai osa laskenta- ja datan tallennus-järjestelmästä sijoitetaan hyvin suojattuun virtuaaliympäristöön, eikä laskentaa tai dataa sallita käytettävän ympäristön ulkopuolella.

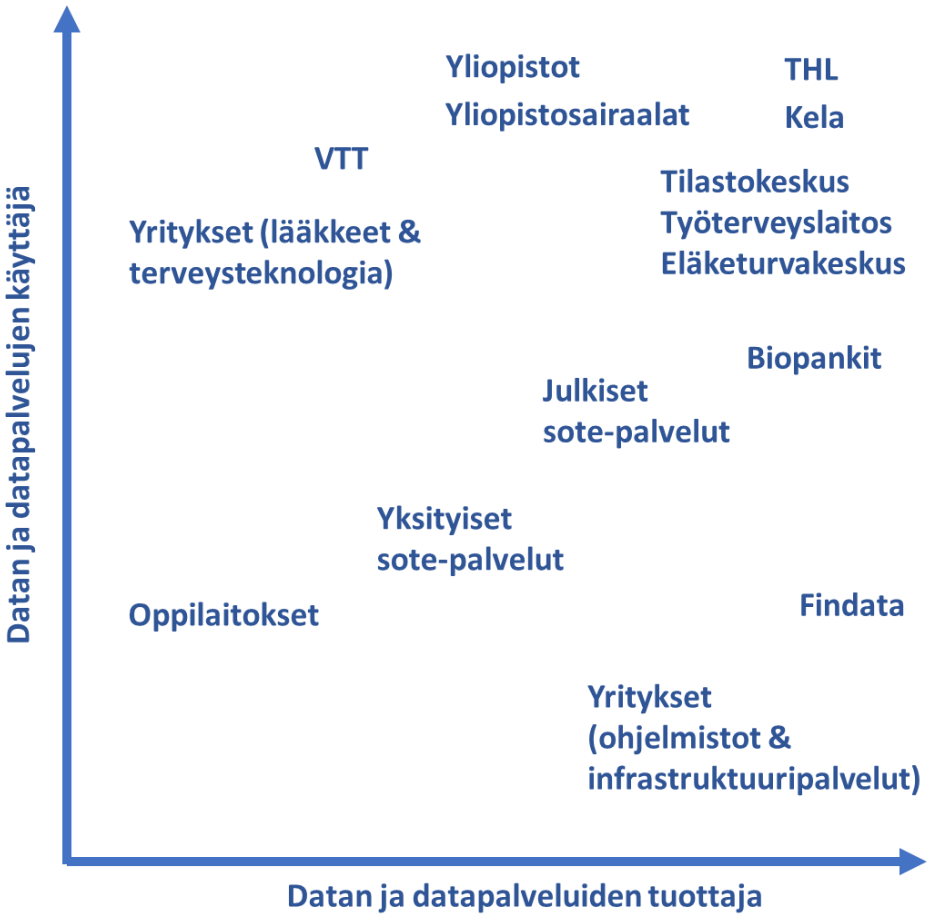
Selvityksessä keskityttiin peilaamaan tätä edellä kuvattua tilannetta nykyhetkeen ja sen tuomiin tulevaisuuden mahdollisuuksiin. Pohjana tarkastelulle oli projektitiimin kokemus datan prosessoinnista tallentamisesta ja käytöstä erilaisissa kansallisissa ja kansainvälisissä projekteissa sekä VTT:n ja THL:n sisäisissä tehtävissä. Lisäksi tiimi tutustui pääpiirteittäin toisiolain esittämiin vaatimuksiin kolmansien osapuolien järjestelmistä sekä keskeisiin EU-tason aktiviteetteihin tulevan 10 vuoden ajalle. Näitä havaintoja peilattiin myös haastatteluissa esiin tulleisiin näkökulmiin.

[11] https://yle.fi/uutiset/3-11605223

4. Tulokset

Haastattelujen kattavuuden hahmottamiseksi haastateltujen tahojen edustamista aloista määriteltiin karkea viitekehys, joka on visualisoitu kuvassa 1. Viitekehyksen perusteella todettiin haastattelujen kattavan kohtalaisen hyvin kohteena olleet aihepiirit ja haastattelujen oletetaan näin edustavan kattavasti erilaisia näkökulmia suhteessa toisiolakiin ja datan toisiokäytön potentiaalisiin sovelluskohteisiin.

Viitekehyksen perusteena oli jako datan tai datapalvelujen tuottajiin ja käyttäjiin. Yrityskentän nähtiin jakautuvan karkeasti kahtia. Haastatelluissa tällaisia olivat lääke- ja terveysteknologian yritykset, jotka käyttävät datapalveluja ja joiden toiminnassa datan käyttö on keskeisessä roolissa. Toinen merkittävä ryhmä oli yritykset, jotka tuottavat datapalveluja ja sinällään mahdollistavat edellä mainittujen toimintaa, mutta datan käyttö itsessään ei ole keskeisessä roolissa yrityksen strategiassa. Esimerkkejä tällaisista ovat ohjelmisto- ja infrastruktuuripalveluja tuottavat ja kehittävät yritykset. Muut haastatellut tahot voidaan nähdä hallinnon ja tutkimusyhteisön edustajina. Erityisesti valtion laitokset, kuten THL ja Kela, tuottavat runsaasti dataa ja datapalveluja sekä myös käyttävät niitä itse tutkimukseen ja tilastointiin. Oppilaitokset toimivat datan parissa, mutta niiden rooli erityisesti datan tuottajana on kuitenkin pieni.



Kuva 1: Haastattelujen toimialojen viitekehys.

4.1 Haastattelujen yhteenveto

4.1.1 Vastaajien taustat

Haastatellut henkilöt edustivat erilaisia organisaatioita pienistä muutaman kymmenen hengen toimijoista tuhansia työntekijöitä työllistäviin organisaatioihin. Vain muutamilla organisaatioilla on erillisiä datan hankintaan erikoistuneita henkilö- tai budjettiresursseja, vaikka haastateltujen edustamat yksiköt tekisivätkin tekoälyyn liittyvää tutkimusta. Tyypillisesti datan hankinnan katsotaan budjetoinnissa olevan osa tutkimustyötä.

Vastaajista osan organisaatioilla on erikseen nimettyjä henkilöitä jotka huolehtivat projektin vaatimista laki- ja lupa-asioista, toisilla taas muutamat erikoistuneet tutkijat ovat kokemuksen kautta keskittyneet näiden hoitamiseen. Usealla toimijalla laki- ja lupa-asiat ovat kuitenkin edelleen tutkimushenkilökunnan vastuulla. Käytännössä yritykset käyttävät useammin myös ulkopuolisia sopimus- ja lupa-asiantuntijoita projektien valmisteluissa, kun taas tutkimusorganisaatioissa lupaprosessit jäävät usein tutkijan vastuulle. Isommissa organisaatioissa sopimuksiin on saatavilla apua erilliseltä lakiosastolta, pienemmät taas saattavat käyttää näissä tilanteissa ulkopuolisia konsultteja ja lakitoimistoja.

Haastatelluilla oli hyvin vaihtelevia henkilökohtaisia taustoja: osalla oli suoraan tekoälyyn liittyvää teknistä tai luonnontieteellistä taustaa, osa taas oli ajautunut muiden tehtävien kautta tekoälyn ja data-analytiikan pariin ja opetellut aihetta työn parissa. Osa haastateltavista käsittelee itse dataa, osa taas kertoi olevansa enemmän koordinoivassa roolissa. On myös huomattava, että haastateltavaksi valikoitiin tahoja, jotka ovat lähtökohtaisesti tekemisissä tekoälyn ja sote-datan analytiikan parissa, ja kiinnostuneita näistä aihealueista. Tutkimuslaitokset ja yliopistot tekevät yhteistyötä sekä julkisen sektorin toimijoiden, että yritysmaailman kanssa. Näitä esimerkkeinä mainittiin yhteistyökumppaneina mm. biopankit ja Cleverhealth Network.

4.1.2 Menetelmät ja data

Vastaajat käyttivät tyypillisissä projekteissa laaja-alaisesti lääketieteellistä dataa terveysrekistereistä kuvantamisdataan. Kotimaisista datalähteistä eniten mainintoja saivat kansalliset rekisterit (THL, Kela, Tilastokeskus), sairaanhoitopiirit ja biopankit. Monissa projekteissa käytettiin erikseen projektin tarpeisiin kerättyä dataa. Myös avointa dataa haluttiin hyödyntää, mikäli tutkimusasetelma sen mahdollisti. Esimerkkeinä nostetut datalähteet painottuivat terveydenhuoltoon. Sosiaalihuollon tiedot eivät saaneet suoria mainintoja, vaikka luultavasti monen vastaajan tutkimuksissa sosiaalihuolto on myös mukana esimerkiksi kansallisten rekisterien hyödyntämisen kautta.

Vastaajien organisaatiot osallistuivat laaja-alaisesti tekoälyn osa-alueiden tutkimiseen ja kehittämiseen. Mainintoja tutkimusaiheista saivat esimerkiksi päätöksenteon tukimenetelmät, riskimallit, datan ryhmittely/klusterointi, synteettisen datan generointi, perinteiset tilastolliset analyysit; luonnollisen kielen analyysi, perinteiset luokittimet (regressio-, puumallit jne.), kuva-analyysi, prosessianalyysi, chat botit, syväoppiminen, hoitopolkujen ennakointi, datan anonymisointi, AI-ratkaisujen tietosuojan parantaminen, datan automaattinen kuratointi, konenäkö, ja signaalianalyysi.

Haastateltujen organisaatioiden suhteet data-analytiikkaan ja terveysdatan hyödyntämiseen ovat myös hyvin vaihtelevia. Osa haastatelluista organisaatioista vain toimittaa dataa käyttäjille, muttei analysoi sitä itse, kun taas osa vain analysoi, muttei hallinnoi itse mitään dataa. Suurin osa vastaajista oli ensisijaisesti datan käyttäjiä.

4.1.3 Tekoälyn ja data-analytiikan mahdollisuudet

Haastateltavilta kysyttiin tekoälyn ja terveysdatan potentiaalisimmista kehityskohteista nyt ja lähitulevaisuudessa 5-10 vuoden sisällä, olettaen, että dataa on laajasti saatavilla. Maininnat heijastelivat meneillään olevia projekteja, mutta myös vielä toteutumattomia mahdollisuuksia hahmoteltiin. Osa vastauksista lähestyi aihetta enemmän perinteisen data-analytiikan näkökulmasta, mutta suurimmassa osassa mukana oli jonkinlainen koneoppimisen ja tekoälyn menetelmä. Kuten jotkut vastaajat pohtivat, tutkimusmenetelmien jatkuvasti kehittyessä on myös määrittelykysymys, milloin tutkimus käsitetään tekoälytutkimukseksi.

Erilaiset dataan ja AI-menetelmiin perustuvat päätöksenteon tuki- ja suositusjärjestelmät koettiin keskeisiksi kehityskohteiksi kaikkien toimijoiden näkökulmasta. Datan yhdistettävyys, siihen liittyvät teknologiat ja niiden kehittäminen nähtiin olevan keskeisessä roolissa näiden järjestelmien mahdollistajana. Päätöksenteon tukijärjestelmistä (ml. AI-pohjaiset suositusjärjestelmät) terveydenhuollon ammattilaiselle ja potilaalle/kansalaiselle mainittiin seuraavia esimerkkejä:

- Resurssien optimointi (mm. hoitaja voi toteuttaa perinteisesti lääkärille kuuluneita tehtäviä)
- Automaattinen palveluohjaus, palvelutarpeen arviointi
- Elintapaohjaus kansalaiselle
- Henkilökohtaiset riskiennusteet (kansalaisen motivointi)
- Huippuosaaminen laajemmin käyttöön (tekoäly oppii asiantuntijoilta)

Edellä mainitut kohteet ovat esimerkkejä soveltavan tutkimuksen alueista, jotka pidemmälle kehitettyinä ratkaisuina voivat tulla varsinaisen sote-palvelutuotannon käyttöön. Selvityksen kysymykset painottuivat enemmän tieteelliseen tutkimukseen kuin varsinaiseen sosiaali- ja terveydenhuollossa tapahtuvaan asiakastietojen ensisijaiseen käyttöön ja tiedolla johtamiseen. Vastaajat mainitsivat kuitenkin myös suoraan tietojohtamisen sovelluskohteita, kuten datan käyttö sote-palvelujen mitoituksessa, vaikuttavuuden mittaaminen/seuranta, prosessien ohjaaminen ja suunnittelu sekä uusien tietolähteiden hyödyntäminen (esim. OmaOlo-palvelun[12] data).

On tärkeää huomata ettei ”tekoäly” ole yksittäinen yleispätevä työkalu kaiken ratkaisuun, ja monet sovelluskohteet ovat tällä hetkellä rajallisia ja lähinnä tehostavat ammattilaisten työtä. Esimerkkinä syväoppimisen (deep learning) menetelmiä voidaan nyt ja lähitulevaisuudessa hyödyntää rajoitetuissa analytiikan kohteissa, kuten lääketieteellisten kuvien analysoinnissa, jossa tulokset voidaan helposti vahvistaa kliinikoiden toimesta.

Tekoälyä voidaan hyödyntää palvelujen automatisoinnissa myös esimerkiksi erilaisten chat-bottien kehittämisessä.[13] On kuitenkin huomattava että chat-bottien kehittäminen ei yleensä välttämättä edellytä yksilötason sote-datan käyttämistä tekoälyn opetusaineistona.

Terveydenhuollon asiakasryhmien tunnistaminen datan avulla ja sen avulla resurssien ja hoidon kohdentaminen sai useita mainintoja. Ilman tutkimusasetelmaa kertyvän yksilötason tiedon hyödyntäminen (real world data) nähtiin keskeisenä trendinä. Real world data viittaa mm. terveydenhuollon tietojärjestelmiin ja kansallisiin rekistereihin kertyvään tietoon sekä potilaan itse keräämää dataan. Näiden tietolähteiden yhdistämistä ja analytiikkaa pidettiin tärkeänä ennakoivan hoidon sekä henkilökohtaisen hoidon kehittämiseksi.

Real world datan analysointia pidettiin hyödyllisenä myös lääkkeiden, hoitomenetelmien ja lääkintälaitteiden kehitystyössä. Tällä alueella hyödyntämiskohteina ovat mm. post-market -tutkimukset, vaikuttavuuden mittaaminen, sekä toimivien lääkemolekyylien löytäminen ja toimimattomien aikainen karsiminen. Terveydenhoidon prosessien ja hoitopolkujen

[12] https://www.omaolo.fi/
[13] Esim. Kelan chattirobotti http://chattirobotti.kela.fi/

ymmärtäminen datan avulla liittyy myös tuotekehitysprosesseihin, ja aiemmin mainitut päätöksenteon tuki ja diagnostiikka voivat olla myös integroituina valmiisiin kaupallisiin tuotteisiin.

Tuotekehityksen lisäksi hyödyt voivat palvella myös suoraan tieteellistä tutkimusta ja väestön terveyden ja hyvinvoinnin seurantaa. Uusilla väestötason seurantamenetelmillä nähtiin myös mahdollisuuksia korvata tutkimushankkeiden puitteissa tehtäviä erilliskyselyitä joko kokonaan, tai paikata kyselyaineistojen kattavuutta.

Datan ja menetelmien sujuvaa liikkumista julkisen ja kaupallisen sektorin, ja toisaalta myös tiedon ensisijaisen ja toissijaisen käyttäjien välillä tulee tukea. Jos terveydenhuollon tietojen ensisijainen ja toissijainen käyttö linkittyisivät entistä vahvemmin, ja toisaalta myös yritysten kaupallisten ideoiden ja akateemisen tutkimuksen välillä olisi enemmän linkkejä, tieteellisen tutkimuksen tuloksia pystyttäisiin nopeasti ja tehokkaasti hyödyntämään terveydenhuollossa ja tuotekehityksessä.

Potilaan itse keräämän datan (ml. monitorointidatan) hyödyntäminen terveydenhoidossa koettiin lähitulevaisuudessa mahdollisuutena esim. pitkäaikaissairauksien kuten diabeteksen hoidossa. Dataa voidaan hyödyntää sairauksien hoidon lisäksi ennakoivassa terveydenhoidossa (mm. sydän- ja verisuonitautien riskin alentaminen, elintapojen seuranta) ja riskipotilaiden etäseurannassa. Ennakoivan terveydenhoidon kokonaisvaltainen käyttöönotto vaatii kustannustehokkaita menetelmiä ja teknologioita monitorointiin ja itseraportointiin. Parhaimmillaan ratkaisut perustuvat reaaliaikaiseen tietoon.

Biosignaalidataa voidaan lähitulevaisuudessa hyödyntää lääketieteellisten kuvien ja biosignaalien (mm. sydänkäyrät) automaattisessa käsittelyssä. Kehityskohteina ovat muun muassa kohteiden/piirteiden tunnistaminen ja näistä johdetut indikaattorit. Esimerkiksi aivosairauksia voi olla mahdollista diagnosoida tällaisen datan avulla.

Yksilöllistetty terveydenhoito ja lääketiede on ollut viime aikoina monessa yhteydessä mainostettu aihepiiri, johon kohdistuu paljon odotuksia.[14]Haastattelujen perusteella tietyissä yksittäisissä taudeissa yksilölliseen hoitoon tähtäävästä tekoälyavusteisesta mallinnuksesta on jo nyt hyötyä, mutta laajemmin hyödynnettynä yksilölliseen lääketieteeseen vaikuttaisi kuitenkin olevan vielä pidempi matka kuin moneen muuhun tekoälyn käytännön sovellukseen. Yksilöllistetty lääkehoito, geenilääketiede ja harvinaissairauksien hoito ovat tällä hetkellä aktiivisia tutkimuskohteita ja niiden odotetaan tulevan pidemmällä aikaperspektiivillä laajamittaiseen käyttöön terveydenhuollossa. Tulevaisuudessa yksilöllistä lääketiedettä voidaan harjoittaa myös ennaltaehkäisyssä yksilöllisten riskianalyysimallien avulla.

4.1.4 Analytiikkaratkaisujen toteuttamisen haasteet

Monet haastateltavat korostivat, että mainittuihin tavoitteisin pääseminen ei ole yksinkertaista. Matkalla nähtiin olevan monenlaisia haasteita. Datan käsittelyyn liittyviä teknisiä haasteita on käsitelty enemmän luvussa 4.2. Sen lisäksi haasteeksi nostettiin erityisesti nykyinen säädösympäristö ja sen tulkinta järjestelmän eri tasoilla, sekä tietoaineistojen puutteelliset metatiedot.

Yleisesti nykyiset säädökset toisiolain puitteissa tarvittavista luvista ja liitteistä koettiin monimutkaiseksi kokonaisuudeksi erityisesti niiden vaihtelevan tulkinnan takia eri toimijoiden välillä. Haastattelujen perusteella vaaditut luvat ja esimerkiksi tutkimussuunnitelman tai datapyynnön tarkkuus vaihtelevat riippuen siitä, mistä dataa halutaan käyttöön ja miltä taholta tätä kysytään. Tämä voi vähentää yleistä kiinnostusta käyttää toisiolain piirissä olevaa dataa ja hidastaa näin teknologian kehitystä Suomessa. Yleisenä pelkona ilmaistiin, että lupa- ja sopimustekniset tehtävät syövät projektien työskentelyyn varatut resurssit eikä varsinaista tutkimusta saada tehtyä kunnolla. Vastaajat kaipasivat hyvien käytäntöjen sekä datan siirtoihin

[14] Ks. esim. https://stm.fi/yksilollistetty-laaketiede

ja rekisterinpitäjyyteen liittyvien sopimusmallien jakamista ja yhtenäistämistä. Jotkut toimijat kokivat myös GDPR-sääntelyn tuomat mahdolliset suuret sakot haasteena ja pelotteena.

Vastaajat toivoivat selkeitä, yleisesti jaettuja tulkintoja ja esimerkkejä sote-datan toisiokäytöstä erilaisiin tilanteisiin. Tietyillä aihealueilla, kuten lääketieteellisen kuvantamisen parissa, suomalaista laintulkintaa pidettiin huomattavasti kireämpänä kuin kansainvälistä yleistä käytäntöä. Tämä on johtanut siihen, että suomalaisista aineistoista ei olla kiinnostuttu aina edes Suomessa, sillä muista maista vastaavat aineistot saadaan paljon helpommin käyttöön. Tekoälyn kehityksen näkökulmasta on myös tärkeää saada tulkintaa ja tutkimusta siitä, miten yksityisyydensuoja ja tutkimustulosten julkisuus saadaan tasapainoon AI-menetelmien tulosten jakamisessa. Ongelmana pidettiin myös sitä, että nykyisessä henkilötietojen luvituksen ja sääntelyn toimintatavassa tutkijalle muodostuu tunne, että häntä jo lähtökohtaisesti epäillään tekevän jotain epäeettistä tai laitonta. Luottamuksen puute latistaa intoa tehdä tutkimusta sote-aineistoilla ja saa tutkijat siirtymään muiden aineistojen tai aihealueiden pariin.

Nykyistä asetelmaa, jossa iso osa datasta ja luvista haetaan Findatan kautta etäkäyttöympäristöön, myös kritisoitiin yleisellä tasolla sen arvaamattoman aikataulun ja siitä syntyvien sivukulujen takia. Erityisesti Findatan palveluja pidettiin kalliina verrattuna tilanteeseen, jossa tutkijat ovat voineet hyödyntää suoraan omia laskentaresurssejaan kuten laskentaklustereita, tutkimusryhmien erikoistyöasemia tai CSC:n yliopistoille tarjoamia palveluja. Kliinisen puolen tutkimuksessa Findatan prosessia pidettiin myös hitaana ja kalliina verrattuna aiempaan tilanteeseen, jossa tutkivat lääkärit ovat saaneet aineistoja suoraan ammatillisten yhteyksien kautta. Kasvaneet kulut liittyivät siis sekä Findatan lupaprosessiin, että etäkäyttöympäristön käyttöön, josta syntyy yliopistotutkijoille uusia kuluja.

Käytännön vaikutuksena nähtiin ongelmalliseksi se, että nykyisin tutkimussuunnitelmat nojaavat usein budjettinsa suhteen instituutioiden omiin infrastruktuureihin, mutta tässä uudessa tilanteessa datan säilyttämiseen ja prosessointiin pitää varata selkeä infrastruktuuribudjetti. Kaupalliset toimijat kokivat ongelmalliseksi myös sen, ettei Findatan lupien suhteen ole varmuutta siitä, kuinka kauan niiden saamisessa kestää, eikä heidän näkökulmastaan luvan hyväksyntää pysty varmuudella sanomaan ennakkoon. Kestämättömäksi koettiin myös tilanteet, joissa uuden tutkijan lisääminen tutkimukseen voi kestää viikkoja tai kuukausia. Lisäksi tutkimusyhteisön edustajat kokivat, että pienellä budjetilla toimivat tutkijat tai esimerkiksi muun työn ohessa tutkimusta tekevien kliinikoiden (joilla ei ole varsinaista sidonnaisuutta organisaatioiden projekteihin tai rahoitukseen) rekisteridatan hyödyntäminen muuttuu vaikeaksi rahallisten tutkimusresurssien puutteen ja lisääntyneen byrokratian takia. Aiheesta on myös käyty julkista keskustelua loppuvuodesta 2020[15] ja uudelleen tammikuussa 2021[16].

Jotkut haastateltavat epäilivät, ettei Findata pysty nykyisessä toimintamallissa hyödyntämään datan omistajien tietotaitoa datan valmistelussa ja luovutuksissa. Yhtenä rajoitteena nähtiin tekoälytutkimuksen näkökulmasta vanhanaikainen suhtautuminen datan ja sen prosessoinnin vaatimaan resursointiin. Koska Findata on sote-datan aihealueella tärkeä toimija, vastaajat pitivät erittäin tärkeänä, että Findatalla on riittävät resurssit ja osaaminen, jottei lupaviranomaisesta itsestään tule tutkimuksen pullonkaulaa. Toisaalta terveysalan yleisempänä ongelmana tekoälytutkimuksessa ja data-analytiikassa pidettiin terveyden-hoitojärjestelmän hitautta muuttaa toimintaansa ja alan yleistä konservatiivisuutta, sekä eri toimijoiden hyvinkin erilaisia intressejä.

Tutkimusaineistojen uudelleenkäytössä niin sanotut FAIR-periaatteet (Findability, Accessibility, Interoperability, and Reuse) ovat nousseet yleisesti tärkeiksi periaatteiksi.[17]

[15] https://www.laakarilehti.fi/ajassa/ajankohtaista/toisiolaki-torppasi-tutkimusta/
[16] https://www.laakarilehti.fi/ajassa/ajankohtaista/onko-toisiolakiin-suunnitteilla-muutoksia/
[17] Periaatteet on nostettu esille mm. PSI-direktiivin toimeenpanon yhteydessä https://avointiede.fi/fi/ajankohtaista/uudistunut-psi-direktiivi-tuo-uutta-puhtia-saatavuuteen

Selvityksen perusteella akateemiset ja julkishallinnon toimijat tuntevat FAIR-periaatteet, kun taas yritystoimijoille aihe on vieraampi. FAIR-periaatteiden ei koettu sopivan sellaisenaan terveysdatalle, koska alueella tietoturvallisuuden ja tietosuojan vaatimukset ovat niin korkeat Toisaalta vastaajat pitivät datan korkealaatuista kuvailua erityisen tärkeänä, joka edistää FAIR-periaatteiden F-kirjainta eli löydettävyyden toteutumista.

Puutteellinen datan kuvailu nostettiin hyvin kattavasti ongelmaksi. Kansalliset rekisterit koettiin paremmin kuvailluiksi kuin alueelliset terveydenhuollon aineistot, mutta metatietojen puutteet vaivaavat kaikkia aineistoja. Jotta datan käyttö pitkältä aikaväliltä olisi tehokasta, tiedonkeruun historialliset erot eri aikoina olisi huomioitava paremmin metatiedoissa. Tietomallien ja ontologioiden hyödyntämistä pitäisi myös tehostaa.

Joissakin tapauksissa ongelmana ei pidetty pelkästään kuvailua, vaan myös datan laatu ja vähäinen harmonisointi (esim. alueelliset erot Suomen sisällä ja maiden välillä) tekevät tutkimuksesta haastateltujen mukaan toistaiseksi mahdotonta tai hyvin vaikeaa. Datan omistajat eivät myöskään haastattelujen perusteella välttämättä aina tunne omaa dataansa ja lupaavat liikaa toisiokäyttöä ajatellen. Jotta toisiokäytön ympärille muodostuisi aidosti toimiva ekosysteemi, itse tuotteen eli datan laatu pitäisi olla kattavasti parempi.

Haastatteluissa tuotiin esille, että edellä mainittuja ongelmia voitaisiin ratkaista myös osaltaan tutkimusorganisaatioiden sisällä. Projekteissa ja organisaatioissa pitäisi keskittää enemmän resursseja datan valmisteluun ja hallintaan, ja hyvästä datan hallinnasta tulisi palkita. Nykytilanne vaatii tutkijoiden keskuudessa myös asennemuutosta ja sen hyväksymistä, ettei kaikkia datoja enää saa omille tietokoneille. Haastatteluissa esitettiin myös ettei kaikissa tilanteissa myöskään tarvitse välttämättä siirtää dataa, vaan tutkijavierailut organisaatioiden välillä fyysisesti ja virtuaalisesti mahdollistavat yhteistyöprojektit, joissa data pysyy alkuperäisen omistajan hallussa.

On huomattavaa, että kaikki data-analytiikka ja tekoälyn kehittäminen eivät vaadi arkaluonteisten yksilötason sote-tietojen käyttöä. Rekisteritutkimus on perinteisesti pohjautunut yksilötason dataan, mutta joissakin tutkimusasetelmissa data-analytiikka voisi onnistua usein pienillä muutoksilla myös aggregoidulla datalla. Kuitenkin selvityksen vastaajien mielestä yksilötason dataan pääsy on tutkimustyön kannalta yleensä erittäin oleellista. Yksilötason aineistoa pidettiin keskeisenä esimerkiksi yksilöllisen hoidon ja päätöksenteon tukijärjestelmien kehittämisessä, ja tarpeellisena myös monimutkaisten tilastollisten mallien kanssa, sillä olisi vaikea määritellä kunkin tilastollisen mallin vaatimaa karkeistusta ennakkoon. On huomattava myös, että kuva- ja mittausaineistoja ei voi yleensä aggregoida, vaan ne tulkitaan Suomessa aina yksilötason aineistoksi.[18]

Koska aggregoidulla datalla ei voida vastata moniin tutkimuskysymyksiin mutta yksilötason henkilötietojen käyttöön liittyy paljon sääntelyä ja hidasteita, vastaajat ehdottivat yhdeksi tutkimusta nopeuttavaksi ratkaisuksi myös synteettisen datan kehittämistä eli datan omistajat tai Findata voisi muodostaa valmiiksi näköisdatasettejä. Tällainen rakenteeltaan aidon näköinen ja tilastollisilta ominaisuuksiltaan oikeaa tietoaineistoa vastaava aineisto helpottaisi erityisesti tutkimuksen suunnittelua sekä tilastollisten mallien ja tietojärjestelmien testaamista ilman tietosuojariskejä, koska se ei sisältäisi todellista yksilönsuojan vaarantavaa henkilötietoa.

[18] https://stm.fi/-/sosiaali-ja-terveystietojen-tietoturvallinen-kasittely-toisiolakiin-liittyva-ensimmainen-linjauspaperi-julki-

4.2 Tekninen selvitys, yhteenveto

4.2.1 Haasteet ja nykytilanne

Terveys- ja hyvinvointidatan turvalliseen hyödyntämiseen tarkoitettuja etäkäyttö ja laskenta-ympäristöjä ei ole vielä laajasti saatavilla (Q4/2020). Toistaiseksi ainoa Findatan kategorisesti hyväksymä laskentaympäristö Findatan yhdistämälle datalle on Findatan oma etäkäyttöympäristö, jonka CSC toteuttaa omassa turvaluokitellussa ePouta-pilvipalvelussaan[19]. Findatan puolesta aineistoja on luovutettu myös muihin ympäristöihin silloin, kun se on ollut välttämätöntä ja Findata on harkinnan ja selvittämisen jälkeen tullut tulokseen, että kyseisissä tapauksissa ympäristö täyttää tarvittavat tietoturvaan ja tietosuojaan liittyvät vaatimukset. Käytännössä kyse on ollut esimerkiksi ympäristöistä, joissa jo käsitellään arkaluontoista sote-dataa ensisijaisessa käytössä. 1.5.2021 jälkeen tällaiset luovutukset eivät kuitenkaan enää ole mahdollisia, ellei laskentaympäristöä ole auditoitu.

Käyttäjän kannalta etäkäyttöympäristö on virtuaalikone, jonne on luotu turvalliset yhteydet käyttäjälle varmistaen samalla, ettei käyttäjä voi siirtää dataa sieltä pois. Useiden haastattelujen kohdalla todettiin, ettei tämä ratkaisu palvele syväoppimista tai neuroverkkojen kehitystä, koska grafiikkaprosessorien (GPU) kaltaista laajaa rinnakkaislaskentatehoa ei näihin ympäristöihin ole vielä saatavilla. Myös ympäristön yleistä laskentatehoa epäiltiin big data -tutkimuksiin nähden riittämättömäksi. Akateemiset toimijat kokivat käyttöympäristön myös kalliiksi verrattuna toimijoiden omien laskentaresurssien käyttöön (yliopistojen laskentaklusterit ja CSC:n yliopistoille tuottamat laskentaresurssit).

Jo aiemmin keskusteluissa Findatan kanssa on tullut esiin, että yhtenä perusteena käyttää muuta kuin Findatan ympäristöä projektin datan prosessointiin voisi olla Findatan ympäristön tarkoitukseen sopimattomat resurssit[20] suhteessa suunniteltuun analytiikkaan. Lainsäädännössä tästä on todettu että luovutus on mahdollista jos se on tutkimuksen kannalta välttämätöntä, mutta näissä tapauksissa ulkopuolisen ympäristön on täytettävä Findatan erillisen määräyksen[21] vaatimukset kolmansien osapuolien tietoturvallisten etäkäyttö-ympäristöjen suhteen. Määräys tulee voimaan 1.5.2021. Tämä avaa mahdollisuuden tutkimustahoille ja kolmansille osapuolille toteuttaa omia laskentaympäristöjä riittävällä laskentakapasiteetilla. Käytännössä esimerkiksi infrastruktuuritoimijat voisivat tarjota tällaisia ympäristöjä kaupallisina ratkaisuina tutkijoille ja muille Findatan asiakkaille. Aikataulullisesti tämä on kuitenkin haastavaa aluksi, sillä asetuksen voimaantulon kannalta eri tahojen tulisi olla jo valmistelemassa omia ratkaisujaan, jos niiden halutaan olevan käyttökunnossa ja auditoituna, kun määräys tulee voimaan. Toisaalta annetut määräykset jättävät auki sen, millä perusteella tällaista järjestelmää saisi ylipäätään käyttää Findatan oman järjestelmän sijasta.

Haastatteluissa useat sairaanhoitopiirien yhteydessä toimivat tahot kertoivat, että näillä tahoilla on jo suunnitteilla omia laskentaympäristöjä, koska Findatan ratkaisu koetaan kankeaksi, pitkäkestoisessa tutkimuksessa liian kalliiksi ja liian rajoittuneeksi. Julkisesti Tampereen yliopistollinen keskussairaala, Tampereen yliopisto ja Helsingin yliopisto ovat ilmaisseet suunnittelevansa yhdessä Findatan määräyksen mukaista laskentaympäristöä[22]. Lisäksi haastatteluissa osa yrityksistä ilmaisi olevansa kiinnostunut erillisten ratkaisujen toteuttamisesta tai käyttämisestä esimerkiksi FinnGen-hankkeessa käytetyn pilviratkaisuun perustuvan mallin mukaisesti. Useat haastatelluista olivat myös kommentoineet toisiolakia sen valmistelun aikana etäkäyttöympäristön osalta. Muutama haastateltava myös pohti, onko nykyinen etäkäyttöympäristöjen kapasiteettihaarukka määritelty vain tilastotieteen näkökulmasta, unohtaen varsinainen data-analytiikka ja koneoppivat ratkaisut. Haastateltavat

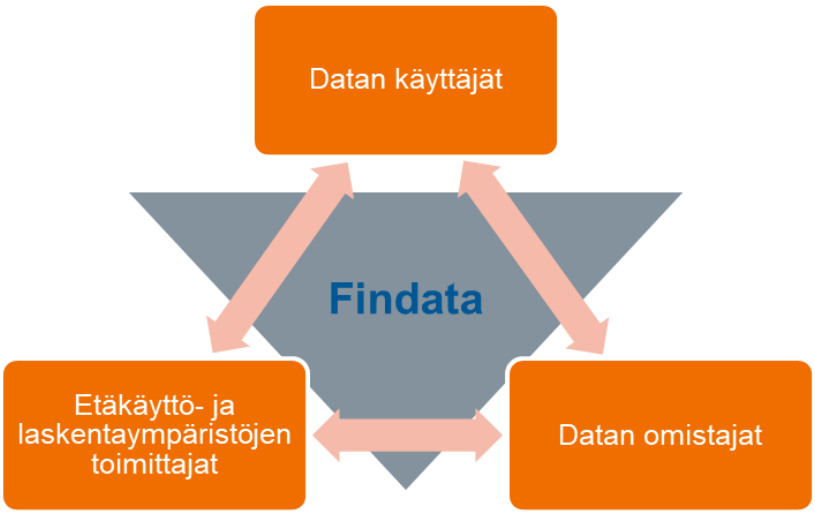
[19] https://research.csc.fi/-/epouta
[20] https://www.findata.fi/palvelut/etakayttoymparisto/
[21] https://www.findata.fi/uploads/2020/10/20ddc0dd-findata-maarays-1-2020-muiden-palveluntarjoajien-tietoturvallisille-kayttoymparistoille-asetettavat-vaatimukset.pdf
[22] https://www.tays.fi/download/noname/%7Bb7079eab-54ae-46df-a10c-93d3c98eafdb%7D/400257

kokivat myös epäselväksi, kuinka monta tutkijaa pystyy yhdessä käsittelemään samaa dataa ja tekemään yhteistä analyysiä.

4.2.2 Tulevaisuus ja mahdollisuudet

Modernissa dataan pohjautuvassa yhteiskunnassa tietoturva on keskeisessä roolissa kaikissa yhteyksissä mukaan lukien tutkimus ja tuotekehitys. Korkeatasoinen tietoturva vaatii ammattilaisten toteuttamia ratkaisuja, joita taas saavutetaan laajassa mittakaavassa helpoiten kaupallisten toimijoiden kautta. Nykyisten määräysten kannalta on siis toivottavaa, että datan prosessointiin ja dataan perustuvan tutkimustyön ympärille syntyy ekosysteemi, joka tarjoaa tutkijoille ja teknologian kehittäjille testattuja ja auditoituja turvallisia ratkaisuja riittävällä laskentakapasiteetilla. Datan prosessoinnin haasteet eivät myöskään ole vain Suomen haaste vaan haaste on globaali, joten hyvin toimiva dataekosysteemi olisi myös potentiaalinen vientituote kansainvälisillä markkinoilla. Toisaalta tällainen ekosysteemi voisi houkutella helpommin myös kansainvälisiä toimijoita Suomeen. Terveysdatan osalta Findata asettuu koko tämän asetelman keskelle. Asetelma on esitetty kuvassa 2.

Alkuvaiheessa on oletettavaa, että monet kolmannen osapuolen tietoturvalliset laskenta- ja etäkäyttöympäristöt ovat julkishallinnon ja mahdollisesti akateemisten toimijoiden toteuttamia mutta pitkällä tähtäimellä myös datan omistajien ja isompien kaupallisten datan hyödyntäjien kannattaa harkita oman käyttöympäristön ylläpitoa. Toisaalta haastattelujenkin kautta on tullut selväksi, että dataan liittyvä lupaprosessi tulee olla riittävän nopea ja läpinäkyvä kaikille prosessin osapuolille (datan omistaja, luvan hakija, laskentaympäristön toimittaja). Riittävän nopea lupakäsittely vaatii myös hyvin määritellyn käsittelyprosessin (ml. hyväksyntä) riittävää automatisointia. Toistaiseksi kaikki tiedossa olevat etäkäyttöympäristöt perustuvat CSC:n ePouta-pilvipalveluun, mutta koska määräyksessä todetaan, että datan tulee säilyä EU:n alueella, myös kaupalliset pilvet voivat tarjota alustan ympäristöille, kun tietoturva otetaan oikein huomioon. Esimerkiksi haastatteluissa ilmeni, että jo nykyäänkin osa sairaan-hoitopiirien datasta prosessoidaan Azuren kaupallisessa pilvessä EU:n alueella.



Kuva 2: Dataekosysteemin toimijat.

Datan saavutettavuuden kannalta keskeistä on myös, että tieto tutkimuskäyttöön soveltuvista aineistoista olisi helposti saatavilla. THL:n ISAACUS-hankkeessa kehitetyt aineistokatalogi[23] ja aineistoeditori[24] ovat hyvä lähtökohta tarvittaville palveluille ja Findata on myös panostamassa niiden kehittämiseen. Teknisenä ratkaisuna ne eivät kuitenkaan riitä yksinään,

[23] https://aineistokatalogi.fi
[24] https://aineistoeditori.fi/

vaan datan omistajien pitäisi olla tietoisia datansa laadusta ja ylipäätään siitä, mitä dataa heillä on tarjota datan käyttäjille. Yleisesti tässä haasteena on datan hajanainen rakenne sote-toimijoiden järjestelmissä ja haasteet datan poiminnassa. Findata antaa määräyksen[25] sote-tiedon toisiokäytön aineistokuvauksista 1.2.2021. Yhdessä STM:n tulevan asetuksen[26] kanssa Findatan määräys tulee velvoittamaan datan omistajat kuvailemaan toisiolain alaiset tietoaineistonsa. Käytännössä datan kuvausten, eli ”metadatan”, luomiseen tarvitaan erillisiä resursseja datanomistajille ja tiiviimpää yhteistyötä parhaiden toimintatapojen määrittelyyn. Ideaalitilanteessa tulevaisuudessa käyttäjä voisi linkittää aineistokatalogin kaltaisesta järjestelmästä itse valmistelemansa listan suoraan Findatan hakemukseen projektin tarvitsemista muuttujista. Samalla tieto hakemuksesta voisi mennä alustavana tietona myös datan omistajalle jo ennen Findatan varsinaista käsittelyä. Toisaalta Findata tai datan omistajat, jos kyseessä olisi vain yhden rekisterinpitäjän alle menevä tutkimus, voisivat määritellä vastaavaan järjestelmään eri projekteille luovutettujen tutkimusaineistojen kuvaukset, mikä lisäisi tutkimusten toistettavuutta ja läpinäkyvyyttä.

Euroopan laajuisessa näkökulmassa datan hyödyntämiseen liittyvien teknisten ja käyttöluparatkaisujen pitäisi myös olla yhteensopivia tuleviin Euroopan laajuisiin järjestelmiin ja verkostoihin. Tällaisia ovat esimerkiksi European Health Data Space (EHDS)[27] ja Gaia-X[28] -hankkeiden tuottamat infrastruktuurit ja prosessit. EHDS-aktiviteettien kannalta Suomi on hyvässä asemassa, koska Sitran vetämänä on alkamassa Towards European Health Data Space (TEHDAS)[29] -EU-hanke keväällä 2021, jossa Findata on myös mukana. Hankkeessa määritellään tulevia vaatimuksia ja rajauksia EHDS-infrastruktuurin ja -hallintamallin kehittämiseksi ja siinä on mukana 22 EU-maata ja 4 ulkopuolista maata. Toinen merkittävä projekti on Gaia-X[30], joka tähtää Euroopan laajuiseen datainfrastruktuuriin, joka mahdollistaa maan rajat ylittävän datan hyödyntämisen turvallisesti ja luotettavasti. Parhaillaan Gaia-X -projektin pohjalta ollaan perustamassa voittoa tavoittelematonta yhteisöä nimeltä Gaia-X AISBL, jonka tavoitteena on edesauttaa yhteiseen datainfrastruktuuriin liittyvää kansainvälistä yhteistyötä ja kehittää verkostoja palvelujen tuottamiseksi. Kansainväliseen datainfrastruktuuriin liittyen on myös tärkeää huomioida niin sanotun PSI-direktiivin (2019/1024)[31] (Public Sector Information) uudistuksen tuomat vaatimukset tietojen saatavuudesta[32].

[25] https://www.lausuntopalvelu.fi/FI/Proposal/Participation?proposalId=ad80bbf9-b11f-4dbf-9d33-087483cbe3f0
[26] https://www.lausuntopalvelu.fi/FI/Proposal/Participation?proposalId=e8657504-0910-4f9a-913b-c602ceec3a92
[27] https://ec.europa.eu/health/ehealth/dataspace_en
[28] https://gaiax.fi/
[29] https://projectsites.vtt.fi/sites/premed/files/workshop2020/Premed_workshop_Kalliola_Sitra.pdf
[30] https://www.data-infrastructure.eu/
[31] https://eur-lex.europa.eu/legal-content/FI/TXT/PDF/?uri=CELEX:32019L1024&from=EN
[32] https://avointiede.fi/fi/ajankohtaista/uudistunut-psi-direktiivi-tuo-uutta-puhtia-saatavuuteen

5. Tulosten tarkastelu

Haastattelujen voidaan katsoa kattavan koko kohteena olleen toimijakentän, kuten on nähtävissä kuvasta 1. Datan kannalta voitiin todeta haastattelujen vastausten olleen vahvasti riippuvaisia haastateltavan roolista suhteessa dataan. Datan tuottajien näkemykset keskittyivät enemmän datan eheyteen ja hallintaan, kun taas datan hyödyntäjillä mielenkiinto oli enemmän siinä, millaisilla prosesseilla dataa saa käyttöön, ja miten dataa voi käyttää. Yhteistä molemmille ryhmille oli, että datan saavutettavuutta ja löydettävyyttä pidettiin keskeisenä teknologian kehityksen kannalta. Jos dataa ei saada käyttöön, ja on vaikea tietää mitä dataa voi saada käyttöön, ei se myöskään kiinnosta tutkijoita tai markkinoita.

Datan käytön kannalta koettiin, että yksityisyysvaatimukset olivat vaikeita sovittaa data-analytiikan vaatimuksiin nykyisissä toimintatavoissa. Osa haastatelluista tutkijoista esitti näihin ratkaisuna hajautettuja analyysejä ja menetelmiä, joissa dataa ei siirretä, vaan analytiikka ja tutkijat tulevat datan luo. Modernien teknologioiden avulla tutkijoiden ei tarvitse tulla fyysisesti datan luo vaan ratkaisu voidaan toteuttaa turvallisten etäyhteyksien ja soveltuvien analytiikkajärjestelmien avulla. Mahdollisuutena tutkimuksen nopeuttamiseksi nähtiin myös näköisdatasetit, jotka mahdollistaisivat datajärjestelmien pystyttämisen ja testaamisen jo lupaprosessien aikana, ennen kuin oikea data saadaan käyttöön. Näissä näköisseteissä datan muoto siis vastaisi oikeaa, mutta sisältö olisi satunnaista.

Lupaprosessien ja projektien suunnittelun kannalta keskeiseksi koettiin, että datalle olisi olemassa riittävän tarkat kuvaukset. Käytännössä tämä tarkoittaisi, että datan omistajien pitäisi kuvata, miten ja mistä data on kerätty ja millaisia muuttujia data sisältää. Analytiikan ja ratkaisujen toistettavuuden ja monistettavuuden tehostamiseksi pitäisi myös sopia datan analysoinnissa hyödynnettyjen menetelmien ja ympäristöjen yhtenevästä dokumentoinnista ja näiden tiedottamisesta erityisesti julkisten toimijoiden keskuudessa parhaiden ratkaisujen laajamittaisen hyödyntämisen varmistamiseksi.

Teknologian ja infrastruktuurin näkökulmasta haastatellut näkivät lähitulevaisuudessa monia haasteita, jotka pitää ratkaista ennen kuin datan todellista potentiaalia voidaan hyödyntää. Yksi keskeinen viesti sairaanhoitopiirien suunnalta oli, ettei ICT-alustan toimittajan pitäisi rajoittaa datan omistajan mahdollisuuksia hyödyntää dataa, vaan omistajalla tulisi olla oikeus päättää itse datan käytöstä. Useat kone- ja syväoppimista hyödyntävät tahot totesivat myös, ettei Findatan tietoturvallinen etäkäyttöympäristö sovellu heidän tutkimuksensa alustaksi, koska GPU- ja teholaskennan tuki puuttuvat ympäristöstä. Ongelmaan on oletettavasti tulossa ratkaisuja myöhemmin, kun ympäristö kehittyy. Toisaalta akateemiset tutkijat kokivat Findatan ympäristön myös kalliina, koska yliopistot eivät voi hyödyntää sitä maksutta kuten CSC:n palveluja yleensä.

Haastatteluissa tuotiin esille, että tutkimuslaitosten, yliopistojen, teollisuuden ja julkishallinnon toimijoiden yhteistoimintaa pitäisi havaintojen perusteella parantaa eikä kaupallista maailmaa pitäisi nähdä ongelmana terveysdatan hyödyntämisessä. Nykyinen keskustelu datan hyödyntämisestä keskittyy usein vain julkiseen ja akateemiseen käyttöön ja kaupallisten toimijoiden osuus keskustelussa jää vähäiseksi. Jos datasta halutaan hyödyntää sen täysi potentiaali, kaikki kolme tahoa tulisi huomioida julkisessa keskustelussa datan toisiokäytön ja datapalvelujen osalta. Tämän keskustelun kannalta esitettiin, että nykyisessä mallissa yksilötason datan hyödyntämisen oikeutus pitäisi perustua tieteellisesti oikeiksi osoitettuihin perusteisiin. Esitetty periaate vastaa hengeltään GDPR:n vaatimusta määritellä ennakkoon mihin tarkoitukseen dataa kerätään ja käytetään ja että vain tarpeelliseksi tiedettyä informaatiota kerätään. Jotta dataa voidaan pyytää käyttöön nojaten tutkittuun tietoon, akateemisen tutkimuksen pitäisi tähdätä osaltaan myös käytännön ongelmien ratkaisuun perustutkimuksen ohessa.

Avoimella keskustelulla sekä tiedeyhteisö että julkishallinto voivat hyötyä kaupallisten toimijoiden valmiista ratkaisuista esimerkiksi datan hallinnassa ja valmistelussa ja siten keskittyä oikeisiin haasteisiin manuaalisen käsityön sijaan. Toisaalta tämä tarkoittaa myös, että

kaupallisten toimijoiden tulisi paremmin ymmärtää käyttäjiensä tarpeita ja olla valmiita sovittamaan järjestelmänsä näihin sen sijaan, että vaadittaisiin käyttäjiä muuttamaan radikaalisti omaa toimintatapaansa. Hyvällä yhteistyöllä tieteellisen tutkimuksen tuloksia voidaan hyödyntää tehokkaasti uusien tuotteiden kehitystyössä kansainvälisille markkinoille. Haastatteluissa alleviivattiin, että menestyviä kaupallisia ratkaisuja saadaan usein parhaiten muodostettua tutkitun tiedon pohjalta, joka tähtää oikeaan, todelliseen tarpeeseen.

Haastattelujen sekä teknisen selvityksen perusteella teknologia ja osaaminen tietoaltaiden ja laaturekisterien pystyttämiseen ovat tärkeitä alan kehittymisen ja myös viennin kannalta. Jos Suomi pystyy profiloitumaan aiheessa kansainvälisesti, tämä voi mahdollisesti tuottaa myös uutta ICT-alan vientiä ja vastavuoroisesti tuoda Suomeen kansainvälisiä tutkimus-investointeja. Jo nykyisellään koetaan, että tähän on lähtökohtia, mutta verkostojen ja yhtenevien toimintatapojen puute sekä lupaprosessien hitaus haittaavat kehitystä.

Tekoälyyn ja suurten arkaluonteisten aineistojen analysointiin liittyviä eettisiä kysymyksiä ei tässä yhteydessä käsitelty, eivätkä haastateltavat niitä myöskään nostaneet omatoimisesti esille. Aiheesta on kuitenkin meneillään hankkeita ja tekoälyn etiikasta on järjestetty viime vuosina myös useita seminaareja. Kuvantamisdatan roolia ja nykyistä anonymiteetin tulkintaa, jossa kuvantamisdata tulkitaan sensitiiviseksi henkilödataksi myös itsenäisenä datana muusta informaatiosta irrallaan, kritisoitiin poikkeavana muuhun Eurooppaan nähden. Tämän koettiin haittaavan merkittävästi kuvantamisdatan AI-ratkaisujen kehittämistä yleisesti koko Suomessa. Aiheeseen liittyen toivottiin neurokeskukselle selvempää roolia ja uutta linjausta kuvantamisdatan sensitiivisyydestä.

6. Suositukset

1. Laaja-alaista julkista keskustelua datapalvelujen tuottajien ja käyttäjien kesken

Tarvitaan laaja-alaista keskustelua ja verkostoa tukemaan uusien teknisten ratkaisujen kehittämistä ja niiden käyttöönottoa toisiodatan hyödyntämisessä Findatan vaatimusten mukaisella mallilla. Uudessa toimintamallissa data ei yhdistämisen jälkeen enää liiku muihin järjestelmiin vaan sen sijaan tutkijat ja analytiikka tulevat virtuaalisesti datan luo.

2. Akateemisten toimijoiden huomioiminen etäkäyttöympäristöjen kehityksessä

Pienten akateemisten toimijoiden tilannetta pitää helpottaa. Findatan turvallisen etäkäyttöympäristön rinnalle voisi muodostaa vastaavan CSC:n palvelun tutkijoille, joilla on lupa käyttää maksutta CSC:n palveluita tutkimuslaitoksen affiliaation kautta. Käytännössä tämä mahdollistaisi, että akateemiset tutkijat voisivat jättää lupapyynnön Findatalle ja anoa projektille CSC:ltä laskentakrediittejä tietoturvallista laskentaympäristöä varten. Kun Findata hyväksyisi lupapyynnön, CSC rekisteröisi pyydetyn tietoturvallisen instanssin Findatalle ja Findata toimittaisi datat järjestelmään. Findatan palvelu voisi edelleen säilyä tämän vaihtoehdon rinnalla kaupallisten toimijoiden perusanalytiikan ratkaisuna, koska kaupallinen puoli ei voi käyttää suoraan CSC:n palveluja.

3. Erilaisten datan toisiokäyttöön perustuvien projektien tarpeiden kartoittaminen ja huomioiminen etäkäyttöympäristöjen tarjonnassa

Findatan ja CSC:n tulisi kartoittaa laajemmin toisiolain puitteissa tehtävän data-analytiikan tarpeita, jotta etäkäyttöympäristöä ja siihen liittyviä palveluja voitaisiin kehittää paremmin erilaisille käyttäjäryhmille soveltuviksi. Palveluun pitäisi esimerkiksi tuoda GPU-resursseja syväoppimista ajatellen, mikä nykyisestä Findatan etäkäyttöympäristöjen tarjoamasta puuttuu. Kun kolmansien osapuolien ratkaisuja alkaa syntyä, pitäisi niistä olla saatavilla julkinen listaus sisältäen suppeat kuvaukset ympäristöjen resursseista, ja siitä millaisia analyysejä näissä voidaan toteuttaa.

4. Toisiolain puitteissa tehdyn tutkimuksen läpinäkyvyys; myönnettyjen lupien tutkimustiivistelmät ja aineistokuvaukset julkisiksi

Tutkimuksen läpinäkyvyyden kannalta olisi hyvä, jos Findata ylläpitäisi julkista listausta tutkimuksista, jotka ovat saaneet hyväksynnän datojen käytölle. Listauksessa pitäisi ilmetä ainakin tutkimuksen pääaihe ja tutkimustahon yhteystiedot. Tutkimusten toistettavuuden ja uusien jatkotutkimusten kannalta olisi myös hyödyllistä, jos Findata voisi tallettaa joksikin ennalta sovituksi ajaksi pyydetyn datan odottamaan mahdollista jatkohanketta. Tällaisissa tapauksissa säästettäisiin huomattavasti kustannuksia ja resursseja, jos Findata voisi uudelleen luvittaa jo aiemmin valmistellun datasetin. Tutkimusaineistojen kuvaukset voitaisiin myös tallentaa esimerkiksi THL:n aineistokatalogiin, huomioiden tietosuojan ja yrityssalassapidon näkökohdat.

5. Tulevaisuuden dataekosysteemin yhteensopivuus kansainvälisen ja erityisesti eurooppalaisen kehyksen kanssa

Etäkäyttöympäristöjen kehittämisessä tulee pitää mukana näkökulma kansainvälisten tulevaisuuden ratkaisujen yhteensopivuudesta. Tällaisia ovat esimerkiksi European Health Data Space (TEHDAS) -hanke ja GAIA-X eurooppalaisen datainfrastruktuurin osalta.

7. Yhteenveto

Projektissa selvitettiin suomalaisen terveysdatan toisiokäyttöä nykyisen toisiolain puitteissa ja terveysdatan potentiaalia data-analytiikassa ja tuotekehityksessä. Projektitiimi teki 20 haastattelua, joiden pohjalta tehtiin huomioita terveysdatan toisiokäytön tilanteesta ja potentiaalista. Lisäksi projektiryhmä teki erillisen selvitystyön teknisten ratkaisujen kehittämisestä ja niihin liittyvistä haasteista. Projektin tulosten perusteella terveysdatan toisiokäytöllä nähdään merkittävä potentiaali ja siihen kohdistuu paljon odotuksia tutkimusyhteisössä, valtionhallinnossa ja yrityksissä. Potentiaalin hyödyntäminen mahdollistaisi resurssien säästämisen ja tukisi uutta kansainvälistä vientiä. Toisaalta suurimpana haasteena nähtiin nykyinen lupaprosessien ja aineistopalvelujen hitaus sekä lainsäädännön vaihteleva kotimainen ja kansainvälinen tulkinta. Tämä raportti toimii lähtökohtana sosiaali- ja terveysministeriön tuleville selvityksille sote-tiedon hyödyntämisestä toisiolain puitteissa tiedolla johtamisen ja resursoinnin näkökulmasta.

Liitteet

Liite 1: Haastattelun saatekirje ja kysymykset

Taustakuvaus

Pyydämme sinua osallistumaan haastattelututkimukseen, jossa kartoitetaan tekoälyn mahdollisuuksia terveyden ja hyvinvoinnin tutkimuksessa. THL:n ja VTT:n yhteistyönä tekemä selvitys liittyy Hyvinvoinnin tekoäly ja robotiikka (Hyteairo) -ohjelmaan, jonka tavoitteena on selvittää ja luoda kuvaa tulevaisuuden tavoitetilasta tekoälyn mahdollisuuksista tutkimuksessa ja tiedolla johtamisessa.

Selvityksen pääfokus on potentiaalisten teknologioiden ja ratkaisujen tunnistaminen ja niihin liittyvien rajoitteiden ymmärtäminen, jotta tilannetta voidaan parantaa tulevaisuudessa. Pyrimme keskittymään siihen, miten toimintamalleja, dataa ja ICT-infrastruktuuria pitäisi kehittää mahdollisuuksien toteuttamiseksi, ja nostamme esille mahdollisia nykyisiä rajoitteita ja esteitä. Erityisesti olemme kiinnostuneita ns. toisiolain ja siihen liittyvien määräyksien vaikutuksesta tekoälyteknologioihin liittyvään tutkimustoimintaan, ja samalla kartoitamme näkökulmia uuden lain mukaisten tietoturvallisten laskentaympäristöjen kehittämiseen.

Projekti laatii katsaustyyppisen selvitysraportin tutkimus- ja kehitystarkoituksiin Sosiaali- ja terveysministeriölle. Projekti toteutetaan vuoden 2020 loppuun mennessä ja sen tulokset julkaistaan erillisessä seminaarissa tammi-helmikuussa 2021.

Haastatteluun osallistuminen on vapaaehtoista. Haastattelut tallennetaan, jos annatte siihen suullisen suostumuksen. Haastattelutallenteet ja niistä tehdyt muistiinpanot säilytetään niin, etteivät muut kuin tämän hankkeen projektiryhmään kuuluvat pääse näkemään niitä. Haastateltavat ovat mukana tutkimuksessa anonyymisti ja tulokset raportoidaan siten, etteivät yksittäiset haastateltavat ole tunnistettavissa. Hankkeesta on laadittu tietosuojailmoitus, joka on saatavissa yhdyshenkilöiltä.

Kysymykset

Kysymykset on valittu koko aihepiirin laajuudelta ja ymmärrettävästi kaikki kysymykset eivät ole kaikille relevantteja. Tästä johtuen voitte jättää kysymyksen väliin, jos kysymys ei sovellu toimialaanne tai rooliinne alalla.

Yleiset kysymykset:

1. Kuinka isoa yksikköä / yritystä edustatte, onko teillä erikseen resursseja olemassa tekoälyä ja datan hankintaa ajatellen?
    - Resurssit voivat olla henkilöitä, tutkimusryhmiä, erillinen budjetti jne.

2. Millainen tausta teillä on henkilönä ja organisaationa data-analytiikkaan ja terveysdataan hyödyntämiseen?
    - Onko data-analytiikka ja sen hyödyntämiseen keskittyvät ratkaisut osa toimintaanne jo ennestään vai oletteko vasta siirtymässä hyödyntämään tekoäly- ja muita dataan pohjautuvia ratkaisuja?

3. Millaisena näette oman roolinne (henkilönä ja organisaationa) suhteessa toisiokäyttöön (datan toimittaja, käyttäjä, infra jne.)?
    - sama organisaatio voi olla useassakin roolissa

Tekoälyn ja data-analytiikan tulevaisuus

4. Mitkä näette terveysdatan potentiaalisimmiksi hyödyntämiskohteiksi tulevaisuudessa olettaen, että dataa on laajasti saatavilla (aika-akseli esim. 5v ja 10v)?
    - Kohteita voivat olla niin datan suora tulkinta kuin monimutkaisten koneoppivien järjestelmien kehittäminen.

5. Datan hyödyntämisen haasteet (henkilö/yritys/yhteisö/kansallisella tasolla), mikä mielestänne voi estää ettei tavoitteisiin päästä.
    - Syitä voivat olla mitkä tahansa rajoitteet toteutuksen kannalta lain tulkinnasta teknologian tasoon.

6. Miten datan saatavuus ja laatu vaikuttavat AI kehitystyöhönne?
    - Esimerkiksi aihetta ei viedä eteenpäin, jos dataa ei ole riittävän helposti saatavilla, tai sen laatu tiedetään heikoksi
7. Miten nykytilaa voitaisiin kehittää, jotta datan saatavuus paranisi?

8. Onko organisaationne tutustunut Findatan turvallisten käyttöympäristöjen määrittelyyn ja varautunut sellaisen käyttöön tai oman auditoidun ympäristön kehittämiseen?
    - https://www.findata.fi/palvelut/etakayttoymparisto/

9. Kuinka monta sote-rekisteritietoja hyödyntävää tekoäly/data-analytiikka -projektia organisaatioyksikössäsi on tällä hetkellä suunnitteilla (arvioitu aloitus 1-2 vuoden sisällä)?

Verkostot

10. Ketkä ovat pääyhteistyökumppaninne tekoälyn / data-analytiikan / datan hyödyntämisessä?
    - Yhteistyö voi keskittyä esimerkiksi datan, osaaminen, teknologian, infrastruktuurin jne. hankintaan.

11. Mitä kautta / miltä taholta yleensä hankitte dataa projekteja / kehitystyötä varten.
    - Oma datan keräys, yksityiset toimijat, rekisterit, sairaanhoitopiirit, tilastokeskus, biopankit jne.

Menetelmät ja ympäristöt

12. Minkä tyyppistä dataa yleensä käytätte?
    - Rekisteridataa, erikseen luvalla kerättyä, avointa, omaa?

13. Millaisia AI ratkaisuja yleensä tutkitte/kehitätte?
    - Esimerkiksi datan ryhmittely/klusterointi, keinotekoisen datan luominen, päätöksenteon tukijärjestelmät jne.

14. Kuinka oleellista on päästä hyödyntämään nimenomaan yksilötason terveysdataa (verrattuna aggregoituun dataan)?

15. Millaisia ympäristöjä tyypillisesti käytätte AI ja dataa hyödyntävissä hankkeissa?
    - Pilvipalvelut, paikalliset tietokoneet tai serverit, auditoitu toimittaja, CSC:n palvelut jne.

16. Miten arvioitte AI/ML/DL menetelmien, tulosten tai datan laatua?
    - AI: Tekoäly (Artificial Intelligence), ML: Koneoppiminen (Machine Learning), DL: Syväoppiminen (Deep Learning)

17. Oletteko tietoisia FAIR (Findability, Accessibility, Interoperability, and Reuse) periaatteista ja jos olette, miten näette nämä periaatteet omalla kohdallanne?
    - https://www.fairdata.fi/tietoa-fairdatasta/fair-periaatteet/
    - https://www.go-fair.org/fair-principles/

18. Onko teillä erikseen erikoistuneita henkilöitä huolehtimaan projektin vaatimista laki ja lupa-asioista esimerkiksi datan saannin ja käsittelyn osalta?
    - Esimerkiksi erikoistuneita laki/sopimustekniikan osaajia tai osasto?
    - Paljonko data-analyytikko/tutkija/tuotekehittäjä joutuu osallistumaan lupa ja sopimusasioihin?

Lopuksi

19. Haluatteko nostaa esille jotain mitä tässä ei suoraan kysytty? Vapaat terveiset ja kommentit.
20. Voiko nimen julkaista raportin haastateltujen listassa? (Suorat vastaukset anonymisoidaan)

Liite 2: Kysymysten vastausten yhteenvedot

1. Kuinka isoa yksikköä / yritystä edustatte, onko teillä erikseen resursseja olemassa tekoälyä ja datan hankintaa ajatellen?
    - Haastatellut henkilöt edustivat yrityksiä laajalla otannalla pienistä muutaman kymmenen hengen toimijoista tuhansia työntekijöitä edustaviin organisaatioihin.
    - Yleisesti vaikka haastatellun edustama yksikkö tekisi tekoälyyn liittyvää tutkimusta, vain muutamilla organisaatioilla on erillisiä datan hankintaan erikoistuneita resursseja (henkilö tai raha). Tyypillisesti tämän toiminnan katsotaan budjetoinnissa olevan osa tutkimustyötä.

2. Millainen tausta teillä on henkilönä ja organisaationa data-analytiikkaan ja terveysdataan hyödyntämiseen?
    - Haastatelluilla on hyvin vaihtelevia henkilökohtaisia taustoja: osalla suoraa tekoälyyn liittyvää teknistä tai luonnontieteellistä taustaa, osa taas ajautunut muiden tehtävien kautta tekoälyn pariin ja opetellut aihetta työn parissa.
    - Organisaatiot hyvin erilaisissa suhteissa data-analytiikan parissa ja terveysdatan hyödyntämisessä. Osa haastatelluista organisaatioista vain toimittaa dataa käyttäjille muttei analysoi itse, kun taas osa vain analysoi muttei hallinnoi itse mitään dataa. Valtion toimijoiden parissa laajamittaisen analytiikan tarpeisiin on vasta havahduttu.

3. Millaisena näette oman roolinne (henkilönä ja organisaationa) suhteessa toisiokäyttöön (datan toimittaja, käyttäjä, infra jne)?
    - Vastaajien näkökulmat riippuivat vahvasti tahosta jota he edustivat. Suurin osa vastaajista ensisijaisesti datan käyttäjiä. Kuvan X ryhmittelyn harmaan laatikon edustajat kokivat itsensä eniten datan käyttäjäksi mutta myös osittain datan tuottajaksi riippuen tilanteesta.
    - Julkiset toimijat ja akatemian edustajat kokivat itsensä datan omistajiksi ja toisaalta oman datansa hyödyntäjiksi. Monet vastaajat myös kokivat heidän ylläpitävän analytiikan ja datan hallinnan infrastruktuureja omaan käyttöönsä.
    - Yritystoimijat näkivät itsensä ensi sijaisesti datan käyttäjinä, mutta myös järjestelmiensä kautta infrastruktuurin kehittäjinä ja datan kerääjinä. Tosin datan keräys oli usein jonkun toisen osapuolen lukuun.
    - Verkostotoimijat näkivät itsensä datan käyttäjinä mutta erityisesti datan hyödyntämisen edistämisen näkökulmasta.

4. Mitkä näette terveysdatan potentiaalisimmiksi hyödyntämiskohteiksi tulevaisuudessa olettaen että dataa on laajasti saatavilla (aika-akseli esim. 5v ja 10v)?
    - RWD, päätöksenteon tukimenetelmät, AI tuettu diagnostiikka, datan yhdistely eri lähteistä ja tämän prosessointi koneoppivilla menetelmillä, automaattinen kuvankäsittely, itsemitatun datan hyödyntäminen osana hoitoprosessia kroonisissa sairauksissa
    - terveysdatan hyödyntäminen sähköisissä potilaspalveluissa; tutkimuksen ja julkisen toiminnan tehostaminen ja parantaminen poikkitieteellisellä yhdistetyllä datalla; Biosignaalien hyödyntäminen ja diagnostiikkamallit; Palvelujen käytön seuranta ja tehostaminen; riskiennusteet eri potilasryhmille; yhdistettyyn dataan perustuva päätöksenteon tuki; pitkällä tähtäimellä yksilöllistetty lääketiede/hoito; kuvantamisdatan tehokas analytiikka; tekoälyn tukema yhdistetty

terveyden/sairauden hoito. Tiedonkulku perusterveydenhuollon ja erikoissairaanhoidon välillä, kaupalliset innovaatiot

- deep learning kuvien analytiikassa; datan saaminen uuden järjestelmän hyödyntämisvaiheeseen; RWD hyödyntäminen laaja-alaisesti; yksilöllinen terveydenhoito; terveiden riskiryhmien seuraaminen ja potilaiden tunnistaminen ennen sairastumista; päätöksenteontuki kaikilla tasoilla, AI pohjaiset suositusjärjestelmät; sairauksien evoluution ennustaminen; hoidon vaikuttavuuden mittaaminen potilastasolla ja reaaliaikaisten järjestelmien mahdollistaminen
- ennakoivaa analytiikkaa; riskikartoitukset ja ennusteet; heterogeenisen datan (EHR) hyödyntäminen laaja-alaisesti; RWD datan hyödyntäminen; datan ja seurannan reaaliaikaisuus; Palveluohjaus ja palvelutarpeen arviointi dataan perustuen. Resurssien kohdentaminen. Yksilökohtaisempaa toimintaa: henkilökohtaistettu hoito ja lääkitys; Kuva ja signaalianalyysi jossa rajoitetuissa kohteissa AI tehokas; AI pohjaiset suositus/muistutus järjestelmät kliinikoille, potilaille ja terveille kansalaisille. Geenilääketiede ja kohdennettu lääkintä; päätöksenteon tukijärjestelmät.

5. Datan hyödyntämisen haasteet (henkilö/yritys/yhteisö/kansallisella tasolla), mikä mielestänne voi estää ettei tavoitteisiin päästä.
    - Kaikki: Dataan pääsy (sopimukset, luvat, tekninen); datan laatu ja vähäinen harmonisointi; maiden ja alueiden erot; puutteellinen metadata/datan kuvaukset. GDPR:n tulkinta eri toimijoiden kesken; lupa ja sopimustekniset aiheet syövät projektien resurssit eikä varsinaista tutkimusta saada tehtyä kunnolla; yksityisyyden suojan ongelmat datan ja AI menetelmien tulosten yksityisyys. Datan omistajat eivät tunne omaa dataansa ja lupaavat liikaa; tulevien datan prosessointiympäristöjen rajoitteet; asenteiden muuttaminen uuteen ajatukseen (dataa ei saa omalle koneelle); rajoittuneet tekniset ratkaisut datan turvalliseen analysointiin; Kliinikoiden vapaa-ajalla tekemien tutkimusten vaikeuttaminen (ei suoraa sidosta mihinkään projektiin tai toimijaan); Ei luoteta tutkijoihin; Ensiökäytön ja toisiokäytön yhteistoimintaa ei huomioida. Liiallinen säätely (laskentaympäristöt) estää kehitystyötä; Findatan hinnat verrattuna aiempiin tilanteisiin (apurahatutkijat huonossa asemassa). Toisiolaki epäselvä eikä selvää tulkintaa vielä ole.
    - Terveydenhoitojärjestelmän inertia ja alan konservatiivisuus, erilaiset intressit eri toimijoiden kesken

6. Miten datan saatavuus ja laatu vaikuttavat AI kehitystyöhönne?
    - Suurin rajoite teknologian kehitykselle on dataan pääsyn rajoitteet; Yksityisyysvaatimukset vaikeita; Datan jakaminen eri tahojen kesken vaatii standardointia; Kuvaus/meta-informaation luominen keskeistä jakamisen ja projektien valmistelun kannalta; ICT-alustan tarjoajan ei pitäisi rajoittaa datan omistajan toimintaa datan kanssa; Yliopistot ja tutkimuslaitokset tarvitsevat koneoppimiseen soveltuvia laskenta ympäristöjä (mm. GPU tuki); Tärkeää että tulevat ratkaisut ovat kansainvälisesti yhteensopivia; Tietoaltaat ja laaturekisterit koetaan tärkeiksi tulevien ratkaisujen kannalta. Datan saaminen vie liikaa aikaa.

7. Miten nykytilaa voitaisiin kehittää, jotta datan saatavuus paranisi?
    - Findatan prosesseja pitäisi selventää; isompia laskentakapasiteettejä tarjolle turvalliseen data-analytiikkaan; järjestelmiä jotka tukevat hajautettua analytiikkaa

(federated learning/analytics): data pysyy alkuperäisessä paikassaan ja vain tulokset ja algoritmit liikkuvat.
- Hyvät sopimusmallit, hyvien käytänteiden jakaminen
- Synteettinen data
- Yliopistojen/tutkimuslaitosten omat tietoturvalliset ympäristöt

8. Onko organisaationne tutustunut Findatan turvallisten käyttöympäristöjen määrittelyyn ja varautunut sellaisen käyttöön tai oman auditoidun ympäristön kehittämiseen?
   - useimmat tahot ovat tutustuneet ja ainakin VSSHP ja HUS suunnittelevat omaa nopealla aikataululla
   - Yliopistot tutustuneet mutta oman järjestelmän kehittäminen haasteellista erityisesti rahoituksen kannalta
   - yrityksissä aihe tunnetaan mutta eivät ole toistaiseksi ryhtyneet toimenpiteisiin aiheen takia

9. Kuinka monta sote-rekisteritietoja hyödyntävää tekoäly/data-analytiikka -projektia organisaatioyksikössäsi on tällä hetkellä suunnitteilla (arvioitu aloitus 1-2 vuoden sisällä)?
   - projektit yleisiä tutkimus- ja yliopisto-organisaatioissa sekä yrityksissä. Julkishallinnolla toiminnassa mukana, jos liittyy oman toiminnan kehittämiseen

10. Ketkä ovat pääyhteistyökumppaninne tekoälyn / data-analytiikan / datan hyödyntämisessä?
    - tutkimuslaitokset ja yliopistot tekevät yhteistyötä sekä julkisen sektorin toimijoiden, että yritysmaailman kanssa
    - erikseen mainittiin mm. biopankit ja Cleverhealth Network
    - alustoista mainittiin omat lokaalit infrastruktuurit, CSC ja Azure

11. Mitä kautta / miltä taholta yleensä hankitte dataa projekteja / kehitystyötä varten.
    - yleisesti: avoin data, projektin kliiniset partnerit, projektissa kerätty data
    - kotimaisia: biopankit, kansalliset rekisterit(THL, KELA, Tilastokeskus), sairaanhoitopiirit.
    - KV: UKBiobank, NIH

12. Minkä tyyppistä dataa yleensä käytätte?
    - avointa/muun toimijan luvalla keräämää / itse luvalla keräämää/ erikseen hankittua
    - laaja-alaisesti lääketieteellistä dataa terveysrekistereistä kuvantamisdataan

13. Millaisia AI ratkaisuja yleensä tutkitte/kehitätte?
    - Päätöksenteon tukimenetelmät, riskimallit, datan ryhmittely/klusterointi, synteettisen datan generointi, perinteiset tilastolliset analyysit; luonnollisen kielen analyysi, perinteiset luokittimet (regressio, puumallit jne), kuva-analyysi, prosessi analyysi, chat botit, syväoppiminen, hoitopolkujen ennakointi, datan anonymisointi, AI ratkaisujen tietosuojan parantaminen, datan automattinen kuratointi, konenäkö, signaalianalyysi

14. Kuinka oleellista on päästä hyödyntämään nimenomaan yksilötason terveysdataa (verrattuna aggregoituun dataan)?
    - Vastaajien mielestä yksilötason dataan pääsy on erittäin oleellista

- rekisteritutkimus perinteisesti pohjautunut yksilötason dataan mutta joissakin tutkimusasetelmissa voisi onnistua usein pienillä muutoksilla myös aggregoidulla datalla
- yksilölliseen hoitoon ja päätöksenteon tukijärjestelmien kehittämiseen keskeistä
- tarpeellista monimutkaisten tilastollisten mallien kanssa (vaikea määritellä mallin vaatimaa aggregaatiota ennakkoon
- Ehdotus: voisiko Findatalla olla järjestely jo aiemmin kysytylle datalle pienemmällä kustannuksella (tutkimuksen toistettavuus)
- Kuva-/mittausaineistoja ei voi aggregoida, ovat aina yksilötason aineistoa.

15. Millaisia ympäristöjä tyypillisesti käytätte AI ja dataa hyödyntävissä hankkeissa?
    - Henkilökohtaisia työkoneita, organisaation omia servereitä, datan omistajien servereitä, CSC:n palveluita
    - GPU-docker kontteja, CSC ePouta virtuaalikoneita, Azure kryptattuna järjestelmänä, kaupalliset pilvet anonymisoidulle datalle
    - R, Python, vähenemässä määrin Matlab, SPSS/SAS
    - Normaalilla kuormalla halvempaa rakentaa pitkässä jaksossa oma kuin käyttää pilvipalvelua

16. Miten arvioitte AI/ML/DL menetelmien, tulosten tai datan laatua?
    - Tarkkuus: ROC-käyrä, sensitivity/specificity, ristiinvalidointi, train-test-validation, verrokkiryhmä korrelaatiot, referenssien käyttö, COS validaatio, avoin data referenssinä, manuaalisesti tarkastettu ground truth
    - Laatu: puuttuvat arvot, miten kerätty, edustavuus, syntaksin laadukkuus, rakenteisuus, manuaalisesti tarkastettu ground truth, ammattilaisten annotaatiot
    - Usein tarkkaillaan edelleen manuaalisesti, vähän työkaluja
    - Biopankki ja rekisteridatalle ei tunnustettuja laatumittareita
    - DigiHealthHub -hankkeen datan laatu ja analytiikka manuaali
        i. https://www.oulu.fi/cht-fi/node/200013

17. Oletteko tietoisia FAIR (Findability, Accessibility, Interoperability, and Reuse) periaatteista ja jos olette, miten näette nämä periaatteet omalla kohdallanne?
    - Akateemiset ja julkishallinnon toimijat tuntevat periaatteet, kun taas yritystoimijoille aihe on vieraampi.
    - Yleisesti koetaan etteivät FAIR-periaatteet sovi sellaisenaan terveysdatalle koska tietoturvallisuuden- ja tietosuojan vaatimukset ovat niin suuret
    - toisaalta tunnustetaan, että dataa pitäisi kuvata paremmin joka täyttää F (findable vaatimuksen)
    - mainittu PSI-direktiivi aiheen mukana
        i. https://avointiede.fi/fi/ajankohtaista/uudistunut-psi-direktiivi-tuo-uutta-puhtia-saatavuuteen

18. Onko teillä erikseen erikoistuneita henkilöitä huolehtimaan projektin vaatimista laki ja lupa-asioista esimerkiksi datan saannin ja käsittelyn osalta?
    - Osalla vastaajista on tällaisia henkilöitä, toisilla taas muutamat tutkijat ovat erikoistuneet kokemuksen kautta näiden prosessointiin. Usealla toimijalla nämä ovat edelleen tutkimushenkilökunnan vastuulla
    - Yritykset käyttävät myös ulkopuolisia sopimus- ja lupa-asiantuntijoita
    - Lupaprosessit usein tutkijan vastuulla, kun taas sopimuksiin löytyy lakiosasto

19. Haluatteko nostaa esille jotain mitä tässä ei suoraan kysytty? Vapaat terveiset ja kommentit.
    - Julkishallinto-yritys-akatemia keskustelua terveysdataan pohjautuvan teknologian osalta pitäisi parantaa ja kaikkien roolia selventää
        i. akatemia luo pohjan teknologialle
        ii. julkishallinnolla on tarve
        iii. yritys luo monistettavan ratkaisun jota voidaan myydä myös muualle
    - tietoturvallisen laskentaympäristön toteutustapa pitäisi julkistaa avoimena tietona, jotta eri toimijat voivat panostaa siihen yhdessä (GPL etc.), ei suljettua IPR:ää
    - KV-yhteistyö aiheessa tärkeää, esimerkkinä Sitran TEHDAS-hanke
    - Pitää huolehtia että Findatalla on riittävät resurssit ja osaaminen ettei siitä tule tutkimuksen pullonkaulaa
    - Findatan nykyinen etäkäyttöympäristö liian kankea, pitäisi voida käyttää joustavasti omia softia, saada riittävät laskentaresurssit (GPU:t jne)
    - toisiolaki ja kuvadata hankalassa yhtälössä, tulkinta poikkeaa yleiseurooppalaisesta (kuvadata anonyymiä)
    - Osaamista sekä datan käytössä, että mahdollistavassa teknologiassa pitää lisätä yleisesti
    - datan keruun erot eri aikoina huomioitava paremmin, metatiedossa edelleen puutteita. Tarkemmat dokumentaatiot siitä mitä on poistettu pseudonymisoinnissa.
    - tietomallien ja ontologioiden hyödyntämistä pitäisi tehostaa
    - datan käyttöluvat vs tutkijavierailut datan omistajien tiloihin -> ei ole pakko siirtää dataa minnekään. Federated analytics, mutta tukijat siirtyvät kirjaimellisesti datan luo. Malli ollut käytössä jo pitkään julkishallinnon ja yritysmaailman sekä akatemian välillä.
    - Haasteena myös GDPR-pelote; isot sakot

20. Voiko nimen julkaista raportin haastateltujen listassa? (Suorat vastaukset anonymisoidaan)
    - Lähes kaikki sallivat tämän, osalla lääkeyrityksistä ehto, että vain jos muutkin lääkeyritykset sallivat