Datan harmoniaa ja standardeja: aineistot pitää käsitellä, kuvailla ja tallentaa samalla tavoin

Mittausinstrumenteista saatua geenidataa sekä kliinistä dataa tuotetaan paljon ja helposti.

Tärkeää on kuitenkin päättää hyvissä ajoin, miten ja missä muodossa raakadata tallennetaan ja miten jälkikäsitelty data luokitellaan ja kuvaillaan mittaustapahtuma mukaan lukien. Metadata eli kuvailutieto datasta on tärkeää määritellä täsmälleen samalla tavalla kaikissa tutkimuslaitoksissa ja laboratorioissa ympäri maailmaa. Muutoin datasta ei saada maksimaalista hyötyä tutkimuksessa, koska sitä ei voida yhdistää muualla tuotettuun dataan.

”Jo omassa tutkimusryhmässä voi olla sekasotkua, jos ei ole käytetty esimerkiksi samoja tiedostoja, jolloin ne eivät ole vertailukelpoisia”, professori **Aarno Palotie** Helsingin yliopiston Suomen molekyylilääketieteen instituutista (FIMM) sanoo.

Kansainvälisen yhteistyön ansiosta on jo saatu muutamia standardeja aikaan. VCF (Variant Call Format) määrittelee bioinformatiikassa käytetyn tekstitiedoston, kun geenisekvenssivariaatioita tallennetaan. BAM (Binary Alignment/Map) on puolestaan formaatti, joka voidaan muuttaa luettavaan tekstimuotoon.

> **“Nyt päästään luomaan yli 1000 organisaation kanssa standardien ohella yhteisiä periaatteita, miten dataa käsitellään ja jaetaan.”**

GA4GH (Global Alliance for Genomics and Health) on kansainvälinen vuonna 2013 perustettu allianssi, jossa on mukana yli 500 bioalan, terveydenhuollon ja IT-alan organisaatiota tavoitteenaan luoda standardeja tutkimuskäyttöön jaettavalle datalle. ELIXIR ja GA4GH päättivät aloittaa marraskuussa 2017 yhteistyön. Sopimus antaa ELIXIR -infrastruktuurille mahdollisuuden vaikuttaa kansainvälisten standardien luomisessa. Sopimus liittyy GA4GH Connect -projektiin, jonka tarkoituksena on saada datastandardit käyttöön kliinisessä potilastyössä vuoteen 2022 mennessä.

Aarno Palotie pitää ELIXIR-keskusten ja GA4GH:n yhteistyötä merkittävänä, koska nyt päästään luomaan yli 1000 organisaation kanssa standardien ohella yhteisiä periaatteita, miten dataa käsitellään ja jaetaan.

”ELIXIR -keskukset ovat hyvin verkostoituneita omissa maissaan ja voivat vaikuttaa paikallisiin käytäntöihin.”

Lainsäätäjän ymmärrettävä erilaiset tarpeet datankäytölle

Datan käsittely ja analysoiminen samalla tavalla ja samoin periaattein vaatii vielä työtä. Suomessa on tavoitteena mahdollistaa genomitiedon hyödyntäminen potilasterveydenhuollossa. Tarkoitus on saada aikaan kansallinen genomitietovaranto, jonka ylläpidosta vastaa Suomen Genomikeskus.

”Suomessa yritetään luoda edistyksellisiä lakeja väestön kliinisen tiedon ja genomitiedon yhdistämiseen ja hyödyntämiseen.”

Kliinisen datan ja tutkimuksessa käytettävän genomidatan yhdistämisessä pitää Palotien mielestä ottaa kuitenkin huomioon datan erilaiset käyttötarkoitukset.

”Potilaslähtöinen analyysi pitää olla juuri oikein. Siinä ei siedetä näytesekaannuksia. Datan pitää olla samassa muodossa ja helposti saatavissa, jos sitä käytetään kliiniseen päätöksentekoon. Tieteellisen genomidatan pitää puolestaan olla joustavaa, nopeasti saatavissa sekä erilaisissa tiedostomuodoissa. Vain joustavalla tavoin tutkimus etenee.”

FinnGen-hankkeessa tutkijat ovat joutuneet käsittelemään huomattavan paljon erilaisia sopimuksia. Tietosuojasäädökset edellyttävät äärimmäisen tiukasti ennalta sovittuja protokollia, mikä on Palotien mielestä ristiriidassa tutkimusideologian kanssa.

”Perustutkimus ei vain etene tällä tavalla, siis ulkoa annettujen protokollien mukaan. Tutkimuksessa prosesseja muunnellaan ja sovelletaan sitä mukaan, miten dataa tuotetaan. Kyseessä ovat toisenlaiset tavoitteet kuin kliinisen genomitiedon hyödyntämisessä.”

Koska vaateet ovat erilaisia, Palotien mukaan tarvitaan rinnakkaiset etenemisreitit kahden näin erityyppisen tiedon hyödyntämiseen. Lainsäädännössä pitäisi tarkentaa, miten kliiniseen tarkoitukseen luotua genomidataa voidaan käyttää myös tutkimuksessa.

“Lainsäädännössä pitäisi tarkentaa, miten kliiniseen tarkoitukseen luotua genomidataa voidaan käyttää myös tutkimuksessa.”

”Sinänsä pitäisi päästä yksimielisyyteen myös siitä, miten tutkimuksessa syntynyttä tietoa voidaan joissakin tilanteissa käyttää järkevällä tavalla kliiniseen päätöksentekoon. Nyt tilanne Suomen nykyisessä biopankkilaissa on epäselvä.”

On luonnollisesti tärkeää huolehtia tarvittavasta tietosuojasta, mutta datankäytön liika tai sekava säätely aiheuttaa ongelmia tutkimukseen. Euroopassa säädösympäristö on osittain rikki. Palotie mainitsee esimerkkinä metadatan louhimiseen.

”Tutkija haluaa esimerkiksi tietää, kuinka monta sellaista yksilöä on suomalaisissa biopankeissa, joilla on tietynlainen genotyyppi ja tietty sairaus ja ikä. Ideaalitapauksessa meillä olisi käytössämme portaali, joka antaisi tämän tiedon reaaliajassa. Tutkija ei näe eikä pääse käsiksi yksilökohtaiseen dataan, jonka tietokone käsittelee konepellin alla. Silloin kun hyödynnetään henkilötietoja, jotka EU-alueella on määritelty tiukaksi, lähtökohtana on, että dataa ei saa käsitellä. Poikkeus on tutkimus, mutta tiukasti tulkittuna se voi jopa vaatia joka kerta erillisen lupaprosessin.”

Datasta käyttö on haastavaa, koska säädösympäristöä on tulkittu lähinnä tietosuojan, ei yksiköille saadun terveyshyödyn näkökulmasta.

”Säädösympäristöä tulisi kehittää niin, että eri viranomaiset tulkitsevat lainsäädäntöä samalla tavalla, jolloin ehdottamaani portaalia voi käyttää. Tällaisen portaalin käyttö ei millään tavalla ole mikään tietosuojauhka, kun sen on asianmukaisesti rakennettu”, Palotie sanoo ja huomauttaa, että metadataan liittyvät säädökset ovat esimerkiksi Yhdysvalloissa Eurooppaa väljempiä.

”Lupaprosesseja Euroopassa on luonnehdittu byrokraattiseksi farssiksi. Lupia voidaan joutua odottamaan jopa vuosia”, Palotie huokaa.

Palotie toivoo, että Suomeen syntyvä uusi genomidataan ja rekisteritiedon toisiokäyttöön liittyvä lainsäädäntö nopeuttaa lupaprosesseja ja selkeyttää datankäyttöön liittyviä säädöksiä.

“Tällä hetkellä eri viranomaiset tulkitsevat lakia eri tavoin, mikä on tutkijan oikeusturvalle heikoin asia.”

”Datapolitiikka pitää olla selkeää ja data kaikkien käytössä. Tällä hetkellä eri viranomaiset tulkitsevat lakia eri tavoin, mikä on tutkijan oikeusturvalle heikoin asia. Uudet lait toivottavasti korjaavat tämän tilanteen. Toivottavasti myös Suomen biopankkilain päivittäminen EU:n uuden tietosuoja-asetuksen (GDPR) takia on samassa linjassa muiden uusien lakien kanssa.”

Ari Turunen

LISÄTIETOJA:

FIMM
Suomen molekyylilääketieteen instituutti (FIMM) on kansainvälinen tutkimuslaitos, jonka toiminta keskittyy sairauksien molekyylitason mekanismien selvittämiseen genetiikan ja lääketieteellisen systeemibiologian menetelmin. Tavoitteena on tutkimustiedon siirtäminen terveydenhuollon käyttöön mm. henkilökohtaista lääketiedettä edistämällä. www.fimm.fi

CSC – Tieteen tietotekniikan keskus Oy
on valtion omistama, opetus- ja kulttuuriministeriön hallinnoima, voittoa tavoittelematon osakeyhtiö. CSC ylläpitää ja kehittää valtion omistamaa keskitettyä tietotekniikkainfrastruktuuria.
http://www.csc.fi
https://research.csc.fi/cloud-computing

ELIXIR
rakentaa infrastruktuurin bioalan tutkimuksen tueksi. Se yhdistää 21 Euroopan maan ja Euroopan molekyylibiologian laboratorion EMBL:n johtavat organisaatiot yhteiseksi biologisen informaation infrastruktuuriksi. Sen Suomen keskus on CSC – Tieteen tietotekniikan keskus Oy.
http://www.elixir-finland.org
http://www.elixir-europe.org

SUOMEN ELIXIR
Puh. +358 9 457 2821 ● e-mail: servicedesk@csc.fi
www.elixir-europe.org/about-us/who-we-are/nodes/finland

www.elixir-finland.org

ELIXIR PÄÄMAJA
EMBL-European Bioinformatics Institute
www.elixir-europe.org