Potilasdatan ansiosta saadaan entistä parempia tekoälymalleja

Ilman dataa ja sen uusiokäyttöä lääketieteellinen tutkimus ei edisty. Kerätyn datan ansiosta voidaan luoda hoitopäätösten tueksi tekoälymalleja, jotka nopeuttavat diagnooseja. Uusia datan analyysitekniikoita tulee koko ajan lisää, mutta miten data saataisiin kaikkien tutkijoiden käyttöön?

Suomeen perustettavan Genomikeskuksen yksi vahvuuksista on biopankkien tietokannat. Keskus vastaisi kansallisen genomitietorekisterin kehittämisestä eli keskitetystä geneettisen tiedon tallennuksesta ja hallinnoinnista. Tarkoituksena olisi saada aikaan laadukas suomalaisten geneettistä variaatiota kuvaava tietokanta. Auria biopankin johtaja Lila Kallio uskoo, että biopankkien ja Genomikeskuksen hyvä yhteistyö voi johtaa merkittäviin tuloksiin geenivarianttien seulonnassa.

”Kun Genomikeskus on perustettu ja se aloittaa toimintansa, voitaneen tutkimuksessa tuotettu genomitieto tallentaa myös genomikeskukseen. Genomikeskus voisi sitten analysoida uudelleen sinne talletettua genomidataa vasten kaiken aikaa karttuvaa referenssigenomitietoa. Näin esimerkiksi uusien tunnistettujen kliinisesti merkittävien varianttien seulonta olisi mahdollista jo aiemmin tuotetusta ja tallennetusta datasta,” sanoo **Lila Kallio**.

Vuonna 2013 Suomessa säädettiin laki biopankeista. Laki mahdollisti biopankkien perustamisen. Suomessa on tällä hetkellä 11 biopankkia. Vuonna 2020 biopankkien verkostoon liittyi Arctic Biopankki, joka säilyttää Oulun yliopiston Pohjois-Suomen alueelta keräämiä laajoja väestöaineistoja. Tutkijat voivat Suomessa hyödyntää kaikkien biopankkien aineistoja Fingenious-verkkopalvelun kautta. Fingenious on digitaalinen työkalu, jonka kautta tutkija voi jättää aineiston luovutuspyynnön. Palvelusta vastaa Suomen biopankkien osuuskunta FINBB.

”Biopankit säilyttävät näytteisiin liittyvää dataa tietoturvallisesti. Biopankkien näytteisiin liittyvä tieto on kaikkien tutkijoiden käytettävissä. Tutkijalla tulee olla tutkimussuunnitelma, jonka biopankkien ohjausryhmät tai eettinen toimikunta hyväksyy. Näytteiden ja niihin liittyvän datan saamiseksi tutkimuskäyttöön biopankeilla on valmis prosessi olemassa.”

Suomessa on poikkeuksellisen kattavat ja laadukkaat terveysalan tietovaran-

not. Vuonna 2019 Suomessa tuli voimaan laki terveystietojen toissijaisesta käytöstä. Datan toisiokäyttö tarkoittaa sitä, että sosiaali- ja terveydenhuollon asiakas- ja rekisteritietoja käytetään muussa kuin siinä ensisijaisessa tarkoituksessa, jonka vuoksi ne on alun perin tallennettu. Laki toisiokäytöstä on luonut paineita myös vuonna 2013 säädetyn biopankkilain uudistamiselle. Datan merkitys biolääketieteen tutkimuksessa kasvaa ja lainsäädännön olisi luotava edellytykset sekä tutkimukselle että tarkoituksenmukaiselle tietoturvalle.

Toisiokäyttö luonnollisesti edellyttää, että ihmisistä kerättyjen tietojen hallinnointi on tietoturvallista. Biopankkeihin saatu ja ihmisistä kerättyjen näytteiden tunnistedata suojataan tarkasti.

”Biopankissa näytteistä poistetaan henkilötunnisteet, jotka korvataan pseudonyymikoodilla. Kun näytteitä luovutetaan edelleen tutkimuksiin, korvataan pseudonyymi vielä uudella, tutkimuskohtaisella koodilla. Koodiavain säilytetään biopankissa. Jos alkuperäiseen näytteeseen pitää palata esimerkiksi siitä löytyneen kliinisesti merkittävän tiedon vuoksi, voidaan se tehdä koodiavaimen avulla,” Kallio sanoo.

Koodiavain mahdollistaa datan uusiokäytön ja tutkimuksen tulevaisuudessa.

”Mikäli näyte anonymisoitaisiin eli tehtäisiin täysin tunnisteettomaksi, siihen palaaminen mahdollisten biopankkitutkimuksessa tehtyjen löydösten vuoksi ei olisi mahdollista, eikä siihen jälkeenpäin myöskään voisi liittää enää lisää näytekohtaista tietoa.”

Lila Kallion mukaan näytteen todellinen arvo muodostuu siitä tuotetusta datasta.

”Dataa syntyy diagnostiikan ja hoidon yhteydessä. Myös tutkimuksissa syntyy näytteestä analysoitua tietoa, mikä tulee palauttaa näytteen omistavalle biopankille liitettäväksi näytteeseen. Biopankki hallinnoi tunnistedatan lisäksi näytteeseen liittyvää kliinistä sekä tutkimuksessa tuotettua dataa.”

Toiveena erilaisia suojaustasoja datan käytölle

Datan toisiokäyttöä koskeva laki keskitti lupaprosessin hallinnoinnin uudelle viranomaiselle Findatalle. Ongelmaksi on tullut lupahakemusten ruuhkautuminen. Hakijat ovat kaikki samalla viivalla riippumatta siitä, koskeeko pyyntö pieniä tai äärimmäisen suuria aineistoja.

Aurian tietopalvelujohtaja ja lääketieteellisen matematiikan dosentti **Arho Virkki** tähdentää, että aineistolle on moninaista käyttöä ja siksi käyttötarkoituksen pitäisi myös määrittää datan suojaamisen tason. Datan toisiokäytön tietoturvaharppaus Suomessa oli Virkin mielestä liian iso askel yhdellä kertaa.

”Äärimmäinen suojaaminen huonontaa datan saatavuutta, jolloin tietoturva ei ole optimaalisella tasolle. Minulle optimaalinen tietoturva tarkoittaa, että aineisto on saatavilla ja sitä voidaan hyödyntää lääketieteen kehitykseen, uusien hoitojen suunnitteluun ja hoidollisten prosessien ohjaamiseen. Optimaalista on, että tieto on käytettävissä mutta samalla riittävästi suojattu. Suojaamisen tason pitäisi tulla riskiperusteisuudesta.”

Koska datanhallinta on kiinteä osa lääkärien ja hoitajien ammattia, datan hyödyntämiseen pitäisi Virkin mielestä löytää tasapaino aineiston saatavuuden

ja suojaamisen välillä. Nyt se on heilahtanut toiseen ääripäähän.

”Aineiston käsittely on esimerkiksi osa lääketieteen opiskelijoiden opintoja. Yksi osahan kouluttautumista on, että opiskelijat käyvät läpi operatiiviset järjestelmät ja poimivat itse tietoja oppiakseen.”

Virkin mielestä ongelma on pitkän aikaan ollut tietoarkkitehtuuri. Lääketieteen ja terveydenhuollon defensiivisyyden ja sääntely takia tietoarkkitehtuuri on perinteistä verrattuna esimerkiksi logistiikkaan tai finanssialaan. Sen takia erilaisten tietojärjestelmien integraatio ei ole hyvä.

Virkki toki myöntää, että sairaalat ovat monimutkaisempia paikkoja kuin esimerkiksi logistiikkakeskukset. Logistiikassa paketti menee linjalle ja se kirjataan järjestelmiin, mutta kun potilas tulee sairaalaan, erilaisia kirjauksia ja järjestelmiä on valtava määrä.

Laki datan toisiokäytöstä määrittelee Virkin mukaan kuitenkin liian tarkasti sen, että yksi järjestelmä sopisi kaikille. Virkin mielestä luvan antaja voisi määrittää erilaisia käyttöympäristöjä tutkijoiden tarpeista riippuen.

”Luvanantaja voisi antaa perustasoisen ympäristön, mikä kelpaa yksinkertainen taulukkolaskenta-tyyppiseen data-analyysiin ja jossa olisi käytettävissä tavallisia tilastotieteen ohjelmointikieliä.”

Jos tutkijat taas tarvitsevat oman ympäristön, tutkijoille pitäisi antaa tarkat ohjeet tietoturvasta jaedellyttäätutkijoiden vakuutukset ohjeiden noudattamisesta.

”Tällöin viranomaiset vastaisit tietoturvan varmistamisesta ja tutkijat vastaisivat toiminnastaan tutkimusrekisterin pitäjälle, eli tutkimusta johtavalle kokeneelle tutkijalle, kuten tähänkin asti. Loppupeleissä on tutkijoiden vastuulla varmistaa, että tutkimustulokset ovat oikein, rehellisiä, tieteellisiä ja anonyymejä.”

Suomessa lääketieteen alan ihmisillä on Virkin mukaan korkea ammattiylpeys ja lääketieteellisen aineiston käsittely on ollut tähänkin asti alan tutkijoilla asianmukaisesti hoidettu. Virkin mielestä tietoturvasta voidaan huolehtia luvanvaraisuuden lisäksi koulutuksella. Tietoturva pitäisikin ottaa osaksi lääketieteen opetusta. Virkki käy säännöllisesti puhumassa Turun yliopistossa kliiniset tutkimuksen perusteet -kurssilla tietoalustoista ja tietoturvasta.

Datan toisiokäyttö luo edellytykset tekoälyn hyödyntämiselle lääketieteesssä

Virkin mukaan lakia datan toisiokäytöstä on alettu korjata. Jos säädökset datan toisiokäytöstä saadaan joustavimmiksi ja lupaprosessit nopeutuvat, tarjoaa se monia mahdollisuuksia tekoälytutkimukseen.

”Nyt kun Suomessa sosiaali- ja terveydenhuollon uudistus meni läpi, on hyvät edellytykset yhdistää perusterveydenhoi-

don ja erikoissairaanhoidon potilastiedot eli potilasdataa voidaan tarkastella kokonaisuutena. Se puolestaan antaa mahdollisuuksia kehittää uusia tekoälysovelluksia kliiniselle puolelle. ”

Tekoälymallien algoritmit voivat tehdä tekstipohjaisia analyyseja potilaskertomuksia tai oppia tunnistamaan kuvista piirteitä, joita voidaan hyödyntää diagnooseissa.

”Tekoälyhän on itse asiassa modernia tilastotiedettä, tilastomatematiikan hienostunut sovellus. Tekoälymalleissa hyödynnetään monimutkaisia tilastollisia menetelmiä. Kun puhutaan koneoppimisesta tarkoitetaan tilastollista oppimista. Nykyään voidaan laskea niin tarkkoja tilastomalleja, että se suorastaan tuntuu taikuudelta.”

Tekoälymallit ovat kiinnostaneet Virkkiä pitkään. Omassa väitöskirjatutkimuksessaan hän laati tekoälymallin ihmisen nukkumisen aikaiseen aineenvaihduntaan. Viime aikoina hän on ollut kehittämässä keuhkoveritulpan ennustemallia tutkijoiden kanssa. Mallia käytetään päätöksenteon työkaluna. Keuhkoveritulppa syntyy, kun muualta elimistössä liikkeelle lähtenyt verihyytymä tukkii keuhkoihin johtavan valtimon. Yleisin oire on äkillinen hengenahdistus. Isoissa keuhkoveritulpissa käytetään verihyytymien liuotushoitoa, jolloin laskimoon annetaan pistokselle veren hyytymistä estävää ainetta.

”Jos on epäilys, että päivystykseen tullut potilas on saanut keuhkoveritulpan, on toimittava nopeasti. Kone pystyy nopeasti vilkaisemaan kuvapakan läpi ja neuvomaan radiologia, mitä kohtaa kuvasta kannattaisi katsoa tarkemmin. Sitten päätetään, pitääkö aloittaa liuotus. Jos ei, niin hoitolinja on toinen. Kaikki pitäisi pystyä tekemään alle 10 minuutissa: keuhkojen kuvaus, diagnoosi ja hoidon aloittaminen.”

Virkin mukaan malli keuhkoveritulpasta oli ensimmäinen tieteellinen testi, jossa yritettiin ratkaista vaikeaa ongelmaa hyvin

pienellä määrällä dataa. Laajempi ja tarkempi tekoälymalli on kuitenkin kehitteillä. Tulossa on tieteellisten julkaisujen lisäksi väitöskirjoja.

”Toteutuessaan malli nopeuttaa päätöksentekoa hoitotilanteessa, mutta se auttaa myös laaduntarkkailussa. Voimme esimerkiksi seuloa jälkikäteen tuliko havaittua kaikki pienetkin keuhkoveritulpat.”

Tekoälymallien kehittäminen edellyttää paljon dataa, joilla algoritmeja opetetaan sekä laskentatehoa.

Varsinais-Suomen sairaanhoitopiiri käyttää Suomen ELIXIR-keskuksen CSC:n e-pouta -pilvipalvelua ja sairaanhoitopiiriin on saatu CSC:n laskentaympäristöön dedikoitu 10 gigabitin yhteys. Virkki toivoo tutkijoille parempaa pääsyä ELIXIR-verkostoon.

”Olisi hienoa, jos tutkijoilla olisi mahdollisuus saada kapasiteettia suoraan ELIXIR-infrastruktuurilta käyttöönsä. Tietoaineisto tulisi suoraan ELIXIRin ympäristöön ja ELIXIR pitäisi huolen riittävästä laskentakapasiteetista.”

ELIXIR-infrastuktuurin Suomen toiminnasta vastaa CSC – Tieteen tietotekniikan keskus. CSC hallinnoi resursseja ja palveluja, jotka ovat osa ELIXIRiä, kuten tunnistautumis- ja auktorisointipalvelut (ELIXIR AAI). ELIXIRissä tavoitteena on muodostaa yksi yhteinen, eurooppalainen tutkimusinfrastruktuuri, jonka ansiosta bio- ja terveystieteiden tutkijat voivat aiempaa helpommin löytää, analysoida ja jakaa aineistojaan. Tutkija voi käyttää ELIXIRin tunnistautumis- ja auktorisointipalveluja luodakseen turvallisen analyysiympäristön ja päästäkseen käsiksi pilveen tallennettuihin tutkimusaineistoihin.

Tekstipohjainen tekoälymalli

Lääkärin kirjoittamaa tai sanelemaa tekstiä voidaan hyödyntää tekoälymalleissa, jotka ovat hoitosuositusten ja diagnoosien apuvälineinä. Lausunnoista ja lauseista voidaan rakenteistaa dataa ja opettaa alogoritmi tekemään päätelmiä. Auria biopankin ja Turun yliopistollisen keskussairaalan ja Turun yliopiston hankkeessa tekoäly opetettiin lukemaan lähes 30 000 potilaskertomuksista tupakointia käsitteleviä teitoja. Tutkija **Antti Karlssonin** vetämässä hankkeessa hyödynnettiin kielimallia nimeltä ULMFiT. Malli koulutettiin VSSHP:n analyysikoneilla suomenkielisen Wikipedian tekstimassaa hyödyntäen. Tämän jälkeen mallista koulutettiin luokittelija käyttäen noin 5 000 tupakointiin liittyvän, käsin annotoidun lauseen aineistoa. Nykyään saatavilla on myös kehittyneempiä, valmiiksi esikoulutettuja suomenkielisiä kielimalleja, joista kuuluisin lienee Googlen BERT-malliin perustuva FinBERT. Sen on tuottanut **Filip Ginterin** vetämä Turun yliopiston tutkimusryhmä käyttäen Suomen ELIXIR-keskus CSC:n laskentatehoa.

Tekoälymallin keräämää dataa hyödyntämällä tutkimus osoitti, että tupakoinnin lopettaminen vaikka vasta syövän diagnoosihetkeen saattaa pidentää elinikää huomattavasti.

”Olen varma, että tulevaisuuden potilastietojärjestelmät eivät ole kaavakemaisia alasvetolaatikoineen, vaan nimenomaan proosallista potilaskertomusta tukevia ja siitä tiedot automaattisesti rakenteistavia versioita,” Karlsson sanoo.

”Tämä on työn tehokkuuttakin ajatellen tärkeää. En halua edes ajatella, millaista monimutkaisien asioiden kirjaaminen mahtaa olla kiireisessä lääkärin arjessa.”

Kun louhitaan isoa massaa dataa, säästetään tavattomasti aikaa ja rahaa. Antti Karlssonin kouluttama tekoälymalli analysoi potilastietoa tupakointiin liittyen. Em. tutkimuksessa malli analysoi 30 000 potilaan sairaskertomuksista saatua tekstidataa. Karlssonin mukaan tällaisia malleja käyttämällä saadaan yli 90% tarkkoja analyyseja jopa tunneissa tai minuuteissa. Se on eri asia kuin että manuaalisesti luettaisiin 30 000 potilaan tekstit ja kerättäisiin muuttujat taulukkoon.

”Parhaassa tapauksessa nämä mallit voisivat olla valmiina saatavilla tietoaltaassa ja voisivat rakenteistaa esimerkiksi tätä tupakkatietoa automaattisesti juuri tutkimuskäyttöä varten,” sanoo Karlsson.

Malli ei anna yksittäiselle potilaalle hoito-ohjetta, mutta luo hyvän kokonaiskuvan.

”Uskon, että ainakin aluksi tulevaisuuden automaattiset järjestelmät keräävät pikemminkin raportointiin ja tutkimukseen tärkeää dataa, kun taas todella tärkeät asiat, kuten esimeriksi lääkeannokset tai allergiat täytyy vielä asiantuntijoiden tarkistaa ja syöttää tiedot manuaalisesti.”

Ari Turunen

LISÄTIETOJA:

A. Karlsson et al. (2021):
Impact of deep learning-determined smoking status on mortality of cancer patients: never too late to quit.
Esmo Open Cancer Horizons.
Vol 3. Issue 3.
https://www.esmoopen.com/article/S2059-7029(21)00135-6/fulltext

Auria biopankki www.auria.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