

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 sit-

ten 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 Einge-

nious-verkkopalvelun kautta. Fingenious on digitaalinen työkalu, jonka kautta tutkija voi jättää aineiston luovutuspyynnön. Palvelusta vastaa Suomen biopankkien osuuskunta FINBB.

”Biopankit säilyttävät näytteisiin liittyvää dataa tietoturvasääntöjen mukaisesti. 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 terveystietovarannot.



not. Vuonna 2019 Suomessa tuli voimaan laki terveystietojen toisijaisesta 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 tutkimusmukaiselle 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 pseudonymikoodilla. Kun näytteitä luovutetaan edelleen tutkimuksiin, korvataan pseudonymiympi 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 anonymisoidisiin eli tehtäisiin täysin tunnistettomaksi, 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 käyttöle

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 tietoturvarahppaus Suomessa oli Virkin mielestä liian iso askel yhdellä kertaa.

”Äärimmäinen suojaaminen huonontaa datan saatavuutta, jolloin tietoturva ei ole optimaalisella tasolla. 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 lääkärin ja hoitajien ammattia, datan hyödyntämiseen pitäisi Virkin mielestä löytää tasapaino aineiston saatavuuden

ja suojaamisen välillä. Nyt se on heilah-
tanut toiseen ääripäähän.

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

Virkin mielestä ongelma on pitkän ai-
kaan ollut tietoarkkitehtuuri. Lääketieteen
ja terveydenhuollon defensiivisyyden ja
sääntely takia tietoarkkitehtuuri on perin-
teistä verrattuna esimerkiksi logistiikkaan
tai finanssialaan. Sen takia erilaisten tieto-
järjestelmien integraatio ei ole hyvä.

Virkki toki myöntää, että sairaalat ovat
monimutkaisempia paikkoja kuin esimer-
kiksi logistiikkakeskukset. Logistiikassa
paketti menee linjalle ja se kirjataan jär-
jestelmiin, mutta kun potilas tulee sairaa-
laan, 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 erilai-
sia käyttöympäristöjä tutkijoiden tarpeista
riippuen.

”Luvantantaja voisi antaa perustasoisen
ympäristön, mikä kelpaa yksinkertainen tau-
lukkolaskenta-tyyppiseen data-analyysiin
ja jossa olisi käytettävissä tavallisia tilas-
totieteen ohjelmointikieliä.”

Jos tutkijat taas tarvitsevat oman ympä-
ristön, tutkijoille pitäisi antaa tarkat ohjeet
tietoturvasta ja edellyttää tutkijoiden vakuu-
tukset ohjeiden noudattamisesta.

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

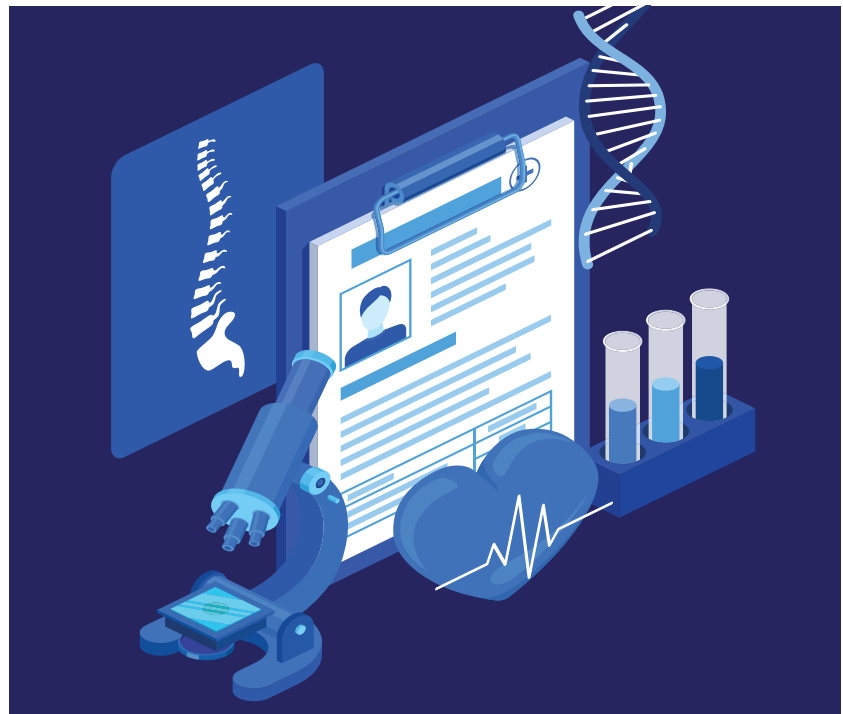
Suomessa lääketieteen alan ihmisillä
on Virkin mukaan korkea ammattitaito
ja lääketieteellisen aineiston käsittely on
ollut tähänkin asti alan tutkijoilla asianmu-
kaisesti hoidettu. Virkin mielestä tietotur-
vasta voidaan huolehtia luvanvaraisuuden
lisäksi koulutuksella. Tietoturva pitäisikin
ottaa osaksi lääketieteen opetusta. Virkki
käy säännöllisesti puhumassa Turun yli-
opistossa kliniset tutkimuksen perusteet
-kurssilla tietoalustoista ja tietoturvasta.

Datan toisiokäyttö luo edellytykset tekoälyn hyödyntämiselle lääke- tieteessä

Virkin mukaan lakia datan toisiokäytöstä
on alettu korjata. Jos säädökset datan toi-
siokäytöstä saadaan joustavimmiksi ja lu-
paprozessit nopeutuvat, tarjoaa se monia
mahdollisuuksia tekoälytutkimukseen.

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

Tekoälymallit ovat kiinnostaneet Virki-
ä pitkään. Omassa väitöskirjatutkimuk-
sessaan hän laati tekoälymallin ihmisen
nukkumisen aikaiseen aineenvaihduntaan.
Viime aikoina hän on ollut kehittämässä
keuhkoveritulpan ennustemallia tutkijoi-
den kanssa. Mallia käytetään päätöksen-
teon työkaluna. Keuhkoveritulppa syntyy,
kun muualta elimistössä liikkeelle lähtenyt
verihyytymä tukkii keuhkoihin johtavan
valtimon. Yleisin oire on äkillinen henge-



don ja erikoissairaanhoidon potilastiedot
eli potilasdataa voidaan tarkastella koko-
naisuutena. Se puolestaan antaa mahdol-
lisuuksia kehittää uusia tekoälysovelluksia
kliiniselle puolelle.”

Tekoälymallien algoritmit voivat tehdä
tekstipohjaisia analyyseja potilaskertomuk-
sia tai oppia tunnistamaan kuvista piirtei-
tä, joita voidaan hyödyntää diagnooseissa.

”Tekoälyhän on itse asiassa modernia
tilastotiedettä, tilastomatematiikan hienos-
tunut sovellus. Tekoälymalleissa hyödynne-
tään monimutkaisia tilastollisia menetelmiä.
Kun puhutaan koneoppimisesta tarkoite-
taan tilastollista oppimista. Nykyään voi-
daan laskea niin tarkkoja tilastomalleja, että
se suorastaan tuntuu taikudelta.”

nahdistus. Isoissa keuhkoveritulpissa käy-
tetään verihyytymien liuotushoitoa, jolloin
laskimoon annetaan pistokselle veren hyy-
tymistä estävää ainetta.

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

Virkin mukaan malli keuhkoveritulpas-
ta 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.

"Toteutuksessaan 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. Tietoaaineisto tulisi suoraan ELIXIRin ympäristöön ja ELIXIR pitäisi huolen riittävästä laskentakapasiteetista."

ELIXIR-infrastruktuurin 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 hoitosuosituksen ja diagnoosien apuvälineinä. Lausunnoista ja lauseista voidaan rakentaa dataa ja opettaa algoritmi tekemään päätelmiä. Auria biopankin 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 syöväen diagnoosihetken saattaa pidentää elinikää huomattavasti.

"Olen varma, että tulevaisuuden potilastietojärjestelmät eivät ole kaavakemai-

sia 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, säädetään 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 analyysieja jopa tunneissa tai minuuteissa. Se on eri asia kuin että manuaalisesti luettaisiin 30 000 potilaan tekstit ja kerättäisiin muututjat taulukkoon.

"Parhaassa tapauksessa nämä mallit voisivat olla valmiina saatavilla tietoaaltaas- ja voisivat rakentaa esimerkiksi tätä tupakkatietoa automaattisesti juuri tutkimuskäyttöä varten," sanoo Karlsson.

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

"Uskon, että ainakin aluksi tulevaisuuden automaattiset järjestelmät keräävät pikemminkin raportointiin ja tutkimukseen tärkeää dataa, kun taas todella tärkeät asiat, kuten esimerkiksi 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](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