

**Kuvantamistutkimuksien lausunnoista
metastaasitietoa tunnistavan kielimallin 1. vaiheen
validointi**

Lääketieteellinen tiedekunta

Kliininen laitos

Syventävien opintojen kirjallinen työ

Laatija:

Johannes Sorvaniemi

9.4.2024

Turku

Turun yliopiston laatujärjestelmän mukaisesti tämän julkaisun alkuperäisyys on tarkastettu
Turnitin OriginalityCheck -järjestelmällä.

Syventävien opintojen kirjallinen työ

Oppiaine: Syöpätaudit ja urologia

Tekijä: Johannes Sorvaniemi

Otsikko: Kuvantamistutkimuksien lausunnoista metastaasitietoa tunnistavan kielimallin 1. vaiheen validointi

Ohjaaja: dosentti Otto Ettala

Sivumäärä: 15 sivua

Päivämäärä: 9.4.2024

Syöpätautien diagnostiikassa, hoidossa ja seurannassa tieto syövän levinneisyydestä on keskeistä. Levinneisyyttä pyritään selvittämään heti syövän diagnoosihetkellä kuvantamistutkimuksin. Kuvantamistutkimuksista saatua tietoa ei kuitenkaan kirjata ja raportoida systemaattisesti, vaan tieto on radiologioiden lausunnoissa vapaamuotoisessa testimuodossa. Syöpätautien tutkimuksen näkökulmasta mielenkiinto kohdistuu lisäksi syövän levinneisyyden tarkkaan aikatauluun ja leviämisen kohde-elimeen. Tieto näistäkin on saatavilla lausuntojen vapaamuotoisesta tekstistä. Tällaisen tiedon automatisoitu löytäminen on haasteellista ja tiedon hakeminen on siksi manuaalista työtä. Tämä on ongelmallista etenkin suurten potilasaineistojen kohdalla.

Ongelman ratkaisemiseksi on kehitetty erilaisia luonnollisen kielen käsittelyyn (natural language processing, NLP) perustuvia tekoälymalleja. Tekoälymallien käytön etuna on automatisoitu tiedonhaku, joka mahdollistaa huomattavasti aiempaa suurempien aineistojen käsittelyn. Kehitetyt mallit ovat pääosin olleet riippuvaisia tietystä syöpätyypistä tai kuvantamismodaliteetista ja kirjallisuuskatsauksen perusteella mallia, joka toimisi suomenkielisille lausunnoille ei ole kehitetty. Tässä tutkimuksessa selvitimme Auria Tietopalveluiden kehittämän NLP-pohjaisen tekoälymallin toimintaa. Tavoitteena oli validoida, miten hyvin mallin 1. vaihe kykenee havaitsemaan lausuntoja, joissa on metastasoinnin kannalta oleellista tietoa.

Mallin toiminnan arvioimiseksi kerättiin 12:sta melanoomapotilaan 101 kuvantamistutkimuksen lausuntoaineisto. Lausuntoaineistosta lääketieteen kandidaatti ja syöpätauteihin erikoistunut lääkäri annotoivat lausunnot virkekohtaisesti metastaattisiin, ei-metastaattisiin tai epäselviin. Manuaalisen annotoinnin tuloksia verrattiin mallin 1. vaiheen kykyyn löytää metastaasitiedon suhteen oleelliset lausunnot.

Tässä tutkimuksessa havaittiin, että käytetyllä lausuntoaineistolla mallin 1. vaihe kykenee lausuntotasolla löytämään 95.8 % (n=23) oleellisista lausunnoista. Ainoastaan yksi lausunto jäi havaitsematta. Virketasolla mallin 1. vaihe sen sijaan löytää vain noin puolet oleellisista lauseista. Rajoittuneen lausuntoaineiston takia mallin toiminnan arvioiminen on kuitenkin epäluotettavaa.

Avainsanat: metastaasi, kuvantaminen, lausunto

Sisällysluettelo

1	Johdanto	4
2	Aineistot ja menetelmät	6
2.1	Lausuntoaineisto	6
2.2	Syövän levinneisyystietoa tekstistä tunnistava malli	6
2.3	Annotointi	7
2.4	Mallin 1. vaiheen toiminnan arviointi	7
3	Tulokset	8
4	Pohdinta	12
4.1	Johtopäätökset	13
	Lähteet	14

1 Johdanto

Syöpätautien diagnostiikassa, hoidossa ja tutkimuksessa tieto syövän levinneisyydestä on keskeistä. Levinneisyys kertoo syöpätaudin ennusteesta ja ohjaa olennaisesti syövän hoitoa. Tieto syövän leviämisestä mahdollistaa hoitojen paremman kohdentamisen ja oikein kohdennetulla hoidolla vältetään altistamasta potilasta tarpeettomalle kärsimykselle ja riskeille, säästetään resursseja ja tehostetaan hoitoa. Syövän leviäminen, leviämisen ajankohta ja kohde-elin ovat täten tärkeitä tietoja paitsi kliinisesti, myös epidemiologisesti ja tutkimuksen näkökulmasta.

Syöpätaudin levinneisyyttä pyritään diagnoosihetkellä ja taudin seurannassa selvittämään kuvantamistutkimuksilla. Nykyään ensisijaisia kuvantamismodaliteetteja ovat tietokonetomografiatutkimukset ja tätä soveltavat menetelmät, kuten aineenvaihduntaa mittaava positoniemissiotomografia (PET). Kuvantamistutkimuksista saatu tieto on vapaassa tekstimuodossa radiologien lausunnoissa. Vaikka lausunnoissa käytetään levinneisyydestä usein toistuvia ilmaisuja, levinneisyyttä ei raportoida tai kirjata systemaattisesti. Kuvantamislöydösten tulkinnanvaraisuus ja levinneisyyden jäsentymätön raportointitapa asettavat haasteita tällaisen tiedon automatisoidulle löytämiselle (Nobel ym. 2020, Pandey ym. 2022). Tiedon hakeminen manuaalisesti on työlästä ja aikaa vievää, mikä on etenkin isojen tutkimusaineistojen kohdalla ongelmallista. Käytössä olevat syöpäkasvainten luokittelutavat, kuten stage ja TNM, eivät sisällä tietoa syövän levinneisyyden kohde-elimestä ja näidenkin luokittelu tapahtuu manuaalisesti.

Tämän ongelman ratkaisemiseksi on kehitetty luonnollisen kielen käsittelyyn (NLP, natural language processing) perustuvia tekoälymalleja (Pons ym. 2016, Casey ym. 2021). Kehitetyt mallit tunnistavat pääosin tekstiä englanniksi, espanjaksi tai kiinaksi ja ne ovat usein kehitetty tiettyä syöpätyyppiä tai kuvantamismodaliteettia ajatellen. (Ahumada ym. 2024, Soysal ym. 2017, Banerjee ym. 2019, Karimi ym. 2021, Groot ym. 2020, Senders ym. 2019, Yang ym. 2022, Park ym. 2022, Pan ym. 2023). Tekoälymallin hyödyntämisen etuna on levinneisyystiedon hankkimisen nopeutuminen. Automatisoitu tiedon louhinta mahdollistaa huomattavasti aiempaa laajempien tietomäärien käsittelyn, joka on avuksi erityisesti retrospektiivisissä tutkimusasetelmissä suurilla potilasaineistoilla. Toisin kuin systemaattisissa raportointitavoissa ja rekistereissä, vapaamuotoisessa tekstissä on kielellisten vivahteiden ja yksityiskohtaisuuden vuoksi mahdollisesti tietoa, jota ei muualta ole saatavilla. Kirjallisuuskatsauksen perusteella vastaavaa mallia, joka toimii suomenkielisille lausunnoille, ja on käytettävissä kaikille syöpätyypeille, ei ole aiemmin kehitetty.

Tämän tutkimuksen tarkoitus on selvittää, kuinka hyvin Auria Tietopalvelujen kehittämän NLP-pohjaisen tekoälymallin 1. vaihe kykenee poimimaan metastasointitietoa sisältäviä lausuntoja kuvantamistutkimuksista.

2 Aineistot ja menetelmät

2.1 Lausuntoaineisto

Mallin ensimmäisen vaiheen validointia varten Auria Tietopalveluiden analyttisestä tietokannasta poimittiin kuvantamistutkimuksien lausuntoaineisto. Analyttinen tietokanta koostuu Uranus -potilastietojärjestelmän datasta, jota ylläpitää 2M IT. Tietokannasta haettiin satunnaistaen melanoomadiagnoosin saaneet potilaat aikavälillä 1.1.2018-15.9.2022. Näiltä potilailta kuvantamistutkimuksien lausuntoja haettiin diagnoosipäivästä 10 vuotta ajassa taaksepäin. Tämän jälkeen potilaat jaettiin menehtyneisiin ja eläviin siten, että kumpaankin ryhmään jäi 50 potilasta. Nämä ryhmät jaettiin edelleen ryhmiin lausuntojen määrän mukaan. Potilaat, joilla oli 0-10 lausuntoa jaettiin ensimmäiseen ryhmään ja potilaat joilla oli 10-20 lausuntoa jaettiin toiseen. Potilaat, joilla oli yli 20 lausuntoa, jätettiin pois lausuntoaineistosta. Jaetuista ryhmistä valittiin satunnaistaen 6 menehtynyttä ja 6 elävää potilasta, joista kummassakin ryhmässä 3 potilaalla lausuntoja oli joko 0-10 tai 10-20 kappaletta. Lopullinen lausuntoaineisto koostui täten 12 potilaasta ja näiden potilaiden 101 lausunnosta aikavälillä 5.7.2009-24.10.2022.

2.2 Syövän levinneisyystietoa tekstistä tunnistava malli

Tässä työssä hyödynnettiin Auria Tietopalvelujen kehittämän syövän metastaasitietoa tekstistä tunnistavan mallin ensimmäistä osaa. Malli on kaksivaiheinen eli ns. hybridimalli.

1. vaihe pyrkii löytämään radiologien lausunnoista virkkeet, joissa esiintyy jokin ennalta määritellyistä kuudesta sanasta (ekspansio, harventuma, leesio, metas, muutos, pesäke). Lista on koottu radiologien suosituksen pohjalta niin, että se kykenee poimimaan mahdollisimman luotettavasti positiiviset virkkeet eli virkkeet, joissa havaitaan syövän etäpesäkkeeseen liittyvää tietoa. 1. vaihe perustuu säännöllisten lausekkeiden hakuun tekstistä eikä sen toiminta perustu kielimalleihin.

Mallin 2. vaiheessa neuroverkko luokittelee mallin ensimmäisessä vaiheessa havaitut virkkeet metastaaseiksi tai ei. Neuroverkko perustuu Googlen kehittämään BERT-kielimalliin ja se on jo aiemmin koulutettu ja validoitu Tyksissä samalle sanalistalle.

Hybridimallin käytön perusteena on mahdollisuus sen eri osien itsenäiseen testaamiseen ja kehittämiseen, jolloin virheiden jäljittäminen on selkeämpää ja mallin osia voidaan validoida toisistaan riippumattomasti. Kehitetyn mallin etuna on sen paikallisuus, joten mallin

käyttö potilastiedon analysoimisessa ei vaadi arkaluontoisen tiedon lähettämistä palvelimille ja mahdollistaa mallin tutkimuskäytön ilman juridisia haasteita.

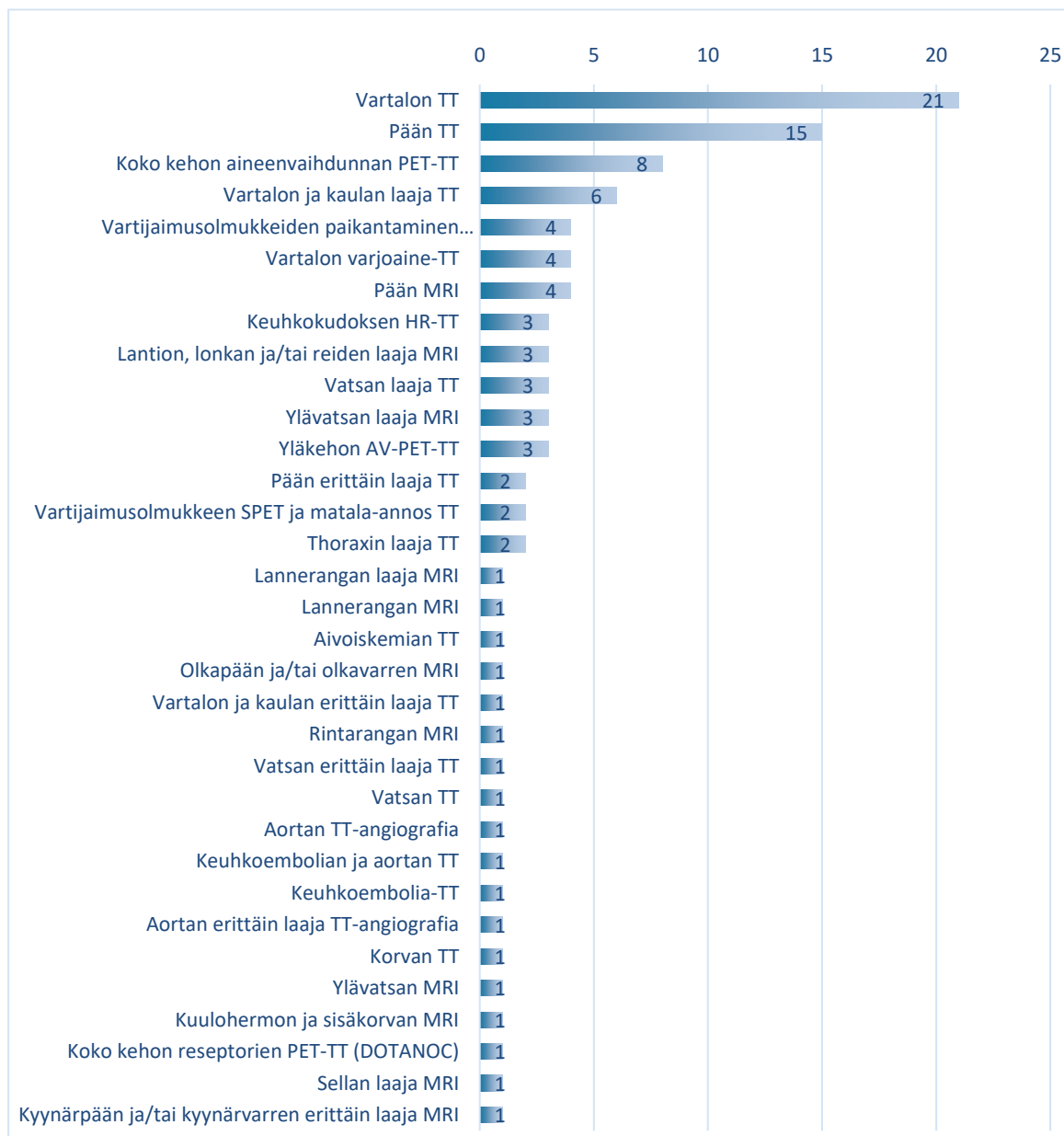
2.3 Annotointi

Mallin 1. vaiheen toimivuuden osoittamiseksi yksi syöpätauteihin erikoistuva lääkäri ja syventäviä opintoja tekevä lääketieteen kandidaatti annotoivat eli luokittelivat lausuntoaineiston virkkeet joko metastaattiseksi, ei-metastaattiseksi tai epäselviksi. Lisäksi metastaasiin viittaava sana tai sanayhdistelmä leimattiin ja liitettiin kohde-elimeen, mikäli tämä oli mahdollista. Negatiivisiksi määriteltiin virkkeet, joissa ei todeta metastasointia ja epäselviksi sellaiset, joissa levinneisyyttä ei voida poissulkea tai selvästi todeta. Lausunnot, joiden ei katsottu sisältävän metastasoinnin kannalta oleellista tietoa, jätettiin annotoimatta. Lausunnon sisällön lisäksi annotoijalla oli tiedossa kuvantamistutkimuksen suoritus-aika ja kutakin potilasta yksilöivä numeraalinen tunniste. Muita potilasta koskevia tietoja, kuten diagnooseja, ei ollut käytettävissä. Annotointi suoritettiin doccano-työkalulla Tyksin tietoverkossa. Tarvittaessa annotoinnin ohjeistuksesta vastasivat tekoälymallin kehittäjänä ja NLP-asiantuntijana toiminut henkilö, yksi syöpätautien erikoislääkäri ja yksi urologian erikoislääkäri. Annotoijat toimivat toisistaan riippumatta.

2.4 Mallin 1. vaiheen toiminnan arviointi

Mallin 1. vaiheen toimintaa arvioitiin tarkastelemalla lukumääriä ja osuuksia manuaalisesti annotoiduista virkkeistä ja näitä verrattiin mallin 1. vaiheen tunnistamiin virkkeisiin. Annotointi ja mallin toiminta tapahtuvat virketasolla, mutta toimintaa tarkasteltiin sekä virketasolla että lausuntotasolla. Tyypillisesti useista virkkeistä koostuvissa lausunnoissa saattoi olla useampia positiivisia, negatiivisia ja epäselviä virkeitä. Jos lausunnossa oli yksikin positiivinen virke, lausunto määriteltiin kokonaisuudessaan positiiviseksi. Lausuntotason arviointi on mallin toiminnan ja soveltamisen kannalta tärkeää, sillä levinneisyystietoa arvioidaan aina koko lausunnon perusteella.

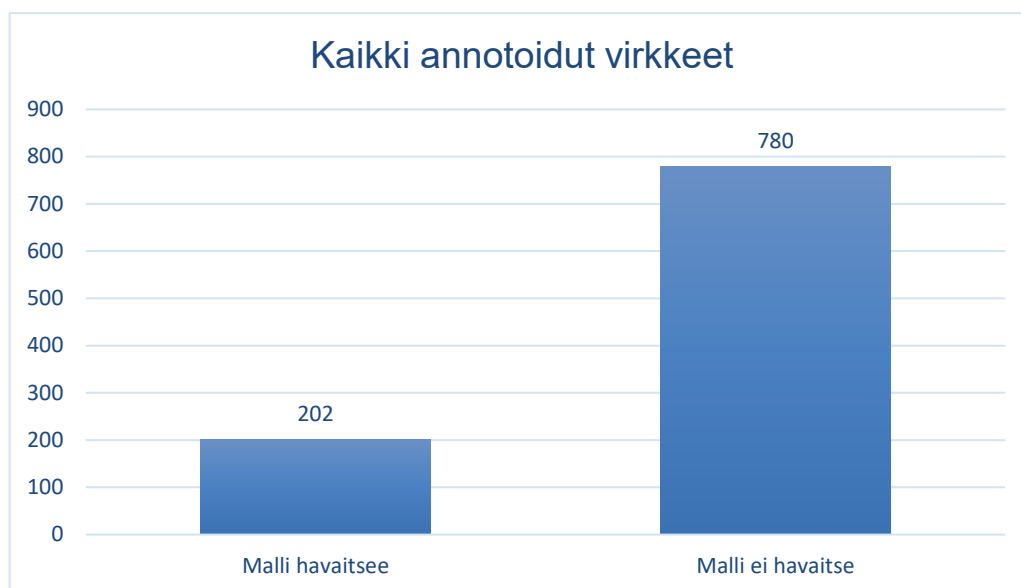
3 Tulokset



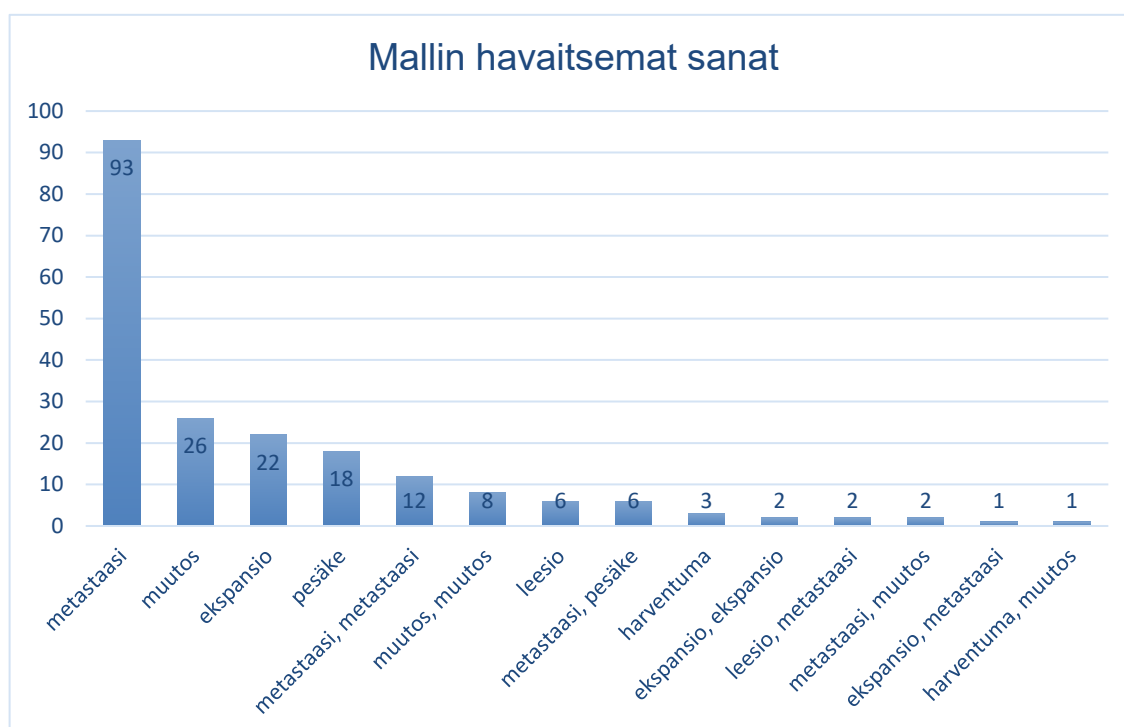
Kuva 1 Annotoinnissa käytetty lausuntoaineisto kuvantamismodaliteetin mukaan

Aineistossa oli 101 lausuntoa, jotka sisälsivät yhteensä 982 virkettä. Loppujen loppuksi 101 lausunnosta annotoitiin manuaalisesti 86. Annotoimatta jätettiin 15 lausuntoa, koska niiden ei katsottu sisältävän metastasointiin liittyvää tietoa. Metastaasitiedon suhteen positiivisia lausuntoja oli 31, negatiivisia 27 ja epäselviä 28. Metastasoinnin suhteen positiiviseksi annotoitiin 80 virkettä. Loput luokiteltiin epäselviksi tai negatiivisiksi.

Kuvassa 1 on listattu työssä käytetyt lausunnot kuvantamismodaliteetin mukaan. Yleisimmät kuvantamismodaliteetit olivat tietokonekuvaus sekä PET:n ja tietokonekuvauksen yhdistelmä.

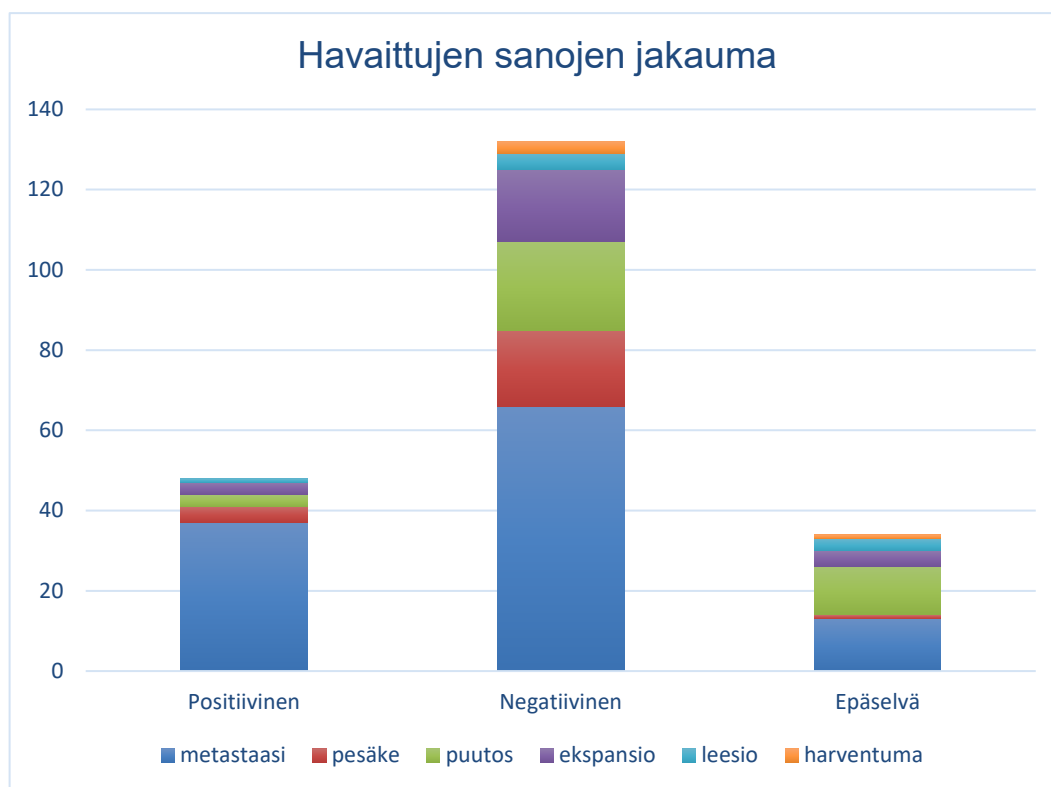


Kuva 2 Mallin 1. vaiheessa havaitsemat ja havaitsematta jääneet virkkeet kaikista annotoiduista virkkeistä



Kuva 3 Mallin 1. vaiheessa havaitsemien sanojen ja sanaparien jakauma

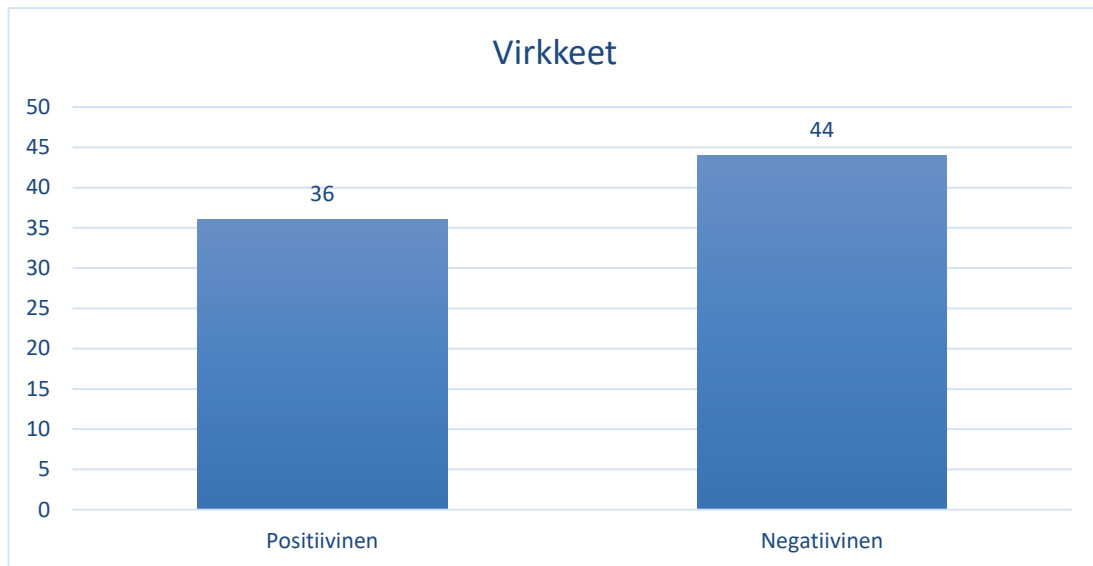
Kuvassa 2 esitetään kuinka suuren osan kaikista 982 manuaalisesti annotoiduista virkkeistä mallin 1. vaihe havaitsee. Kaikista annotoiduista virkkeistä malli 1. vaihe löytää 202 kappaletta. Kuvassa 3 esitetään mallin 1. vaiheen havaitsemat 202 virkettä sen mukaan, mitä sanoja tai sanapareja ne sisältävät. Näistä virkkeistä 93 sisältää sanan metastaasi, 26 sanan muutos, 22 ekspansio ja 18 pesäke. Loput koostuvat sanalistan muista sanoista tai näiden yhdistelmistä.



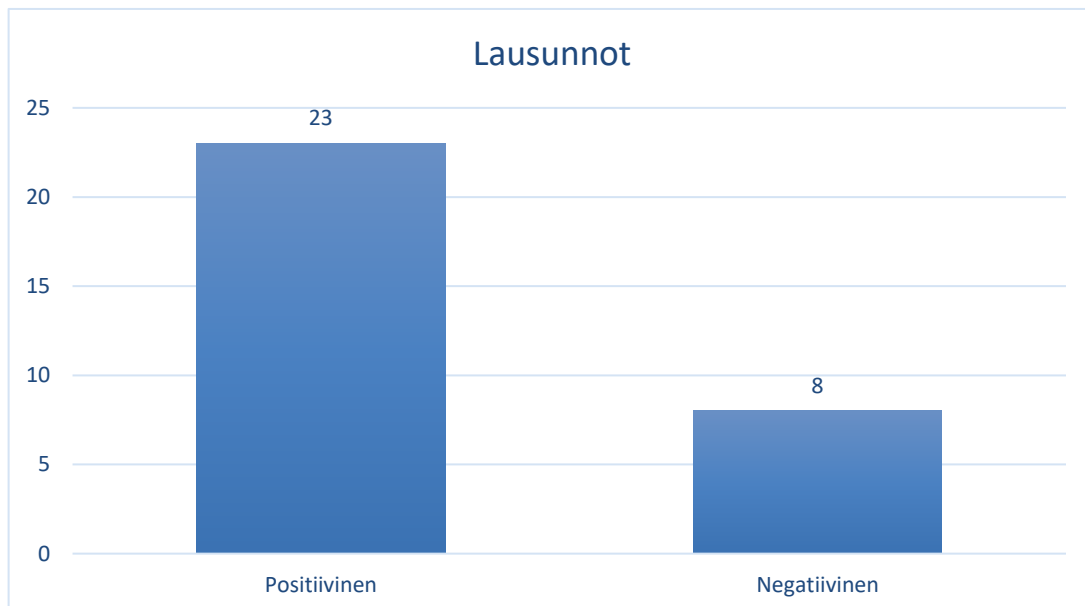
Kuva 4 Mallin 1. vaiheen havaitsemien virkkeiden jakauma näille annotoitujen leimojen ja havaittujen sanojen mukaan

Kuvassa 4 on esitetty mallin 1. vaiheen havaitsemien virkkeiden jakautuminen annotoitujen leimojen mukaan sekä sen mukaan, mitä sanoja ne sisältävät. Mallin 1. vaiheen löytämät 202 virkettä sisältävät manuaalisesti annotoituja positiivisia virkkeitä 48, negatiivisia virkkeitä 132 ja epäselviä virkkeitä 34.

Manuaalisesti positiivisiksi annotoiduista virkkeistä mallin 1. vaihe löytää 44 ja 36 jää havaitsematta (Kuva 5). Lausuntotasolla mallin 1. vaihe löytää positiivisiksi annotoiduista 31 lausunnosta 23 eli 8 positiivista lausuntoa jää havaitsematta (Kuva 6). Havaitsemattomista lausunnoista yksi sisältää termin ”mts-näköistä” kuvaamaan metastaasia, yksi sisältää termin ”melanoomamuutoksen” ja kuusi viimeistä kuvaavat saman potilaan yhtä kasvainta eri termein. Näistä kahdeksasta havaitsemattomasta lausunnosta yksi on todellinen metastaasi ja loput seitsemän ovat primaarikasvaimia. Seitsemän lausuntoa annotoitiin väärin positiivisina. Kun tämä virhe korjataan, todellisten positiivisten lausuntojen määrä on 24 ja näistä malli havaitsee 23 eli ainoastaan yksi jää havaitsematta. Kun annotoinnista johtuvat virheet korjataan, malli tunnistaa 95,8 % metastaasin sisältävistä lausunnoista.



Kuva 5 Positiiviseksi annotoitujen virkkeiden jakauma mallin 1. vaiheen havaitsemisen mukaan



Kuva 6 Positiiviseksi annotoitujen lausuntojen jakauma mallin 1. vaiheen havaitsemisen mukaan

4 Pohdinta

Tässä validointityössä havaittiin, että mallin 1. vaihe kykenee käytetyllä lausuntoaineistolla havaitsemaan oleellista metastaasitietoa sisältävät lausunnot suurella todennäköisyydellä kuuden sanan listaa käyttäen. Kun manuaalisesta annotoinnista johtuneet virheet korjataan, mallin 1. vaihe havaitsee 95,8 % metastaaseja sisältäneistä lausunnoista. Vain yksi todellinen positiivinen lausunto jää havaitsematta ja tässä lausunnossa metastasoinnista käytettiin poikkeuksellista ilmaisua ”mts-näköistä”.

Mallin tärkein ominaisuus on tunnistaa lausunnoista metastaasiin viittaava tieto mahdollisimman herkästi, jotta kaikki metastaaseihin viittaavat lausunnot löydetään. Mallin 2. vaihe, joka tekoälyn avulla määrittelee sille 1. vaiheessa syötetyt virkkeet joko metastaattisiksi tai ei metastaattisiksi on validoitu toimivaksi, mutta 1. vaiheen sanalistan toimivuutta ei ole validoitu. Koko mallin herkkyyden kannalta mallin 1. vaiheen kyky havaita oleelliset virkkeet on olennaista, joten käytetyn sanalistan toimivuuden validointi ja sanalistan kehittäminen on tärkeää.

Tämän työn puitteissa mallin 1. vaihe kykenee lausuntotasolla havaitsemaan metastaasitiedon erittäin hyvin, mutta virketasolla vain noin puolet oleellisista lauseista löytyy. Lausuntotasaisen tiedon tunnistaminen on riittävää, jos halutaan tietää metastasoinnin ajankohta, mutta virketason tunnistamista tarvitaan, jos mallia halutaan selvittää lisäksi metastasoinnin kohde-elin. Virketason alhainen herkkyys tulee todennäköisesti kuitenkin vaikuttamaan myös lausuntotason suorituskyvyssä, jos aineisto on riittävän laaja.

Annotoinnin laadulla on keskeinen merkitys mallin toiminnan arvioimisen luotettavuudelle. Annotoinnissa tehtiin useita virheitä, joiden korjaaminen muuttaa tuloksia ja näiden tulkintaa oleellisesti. Osa virheistä oli inhimillisiä erehdyksiä ja osa taas tietotaidosta johtuvia. Merkittävä virhelähde annotoinnissa oli myös virkkeiden luokittelu kolmeen kategoriaan. Yksiselitteisesti selvää jakoa metastasoinnin ja ei-metastasoinnin välillä on monesti vaikea tehdä. Tässä työssä ainoa mallilta havaitsematta jäänyt positiiviseksi annotoitu sanayhdistelmä ”mts-näköistä” ei sekään ole yksiselitteisesti positiivinen levinneisyyden suhteen. Annotoinnin laatua voitaisiin parantaa laatimalla yhteisiä ohjeistuksia annotoijille sekä tarkastelemalla ongelmalliseksi havaittuja lausuntoja tarkemmin.

Validoinnissa käytetty lausuntoaineisto oli määrältään rajoittunut eikä vastannut sisällöltään yleisimpiä levinneisyystutkimuksia. Yhteensä 101 lausunnosta metastaasitiedon suhteen olennaisia oli 86 ja näistä todellisia positiivisia lausuntoja oli 24. Alhainen positiivisten lausuntojen määrä tekee mallin 1. vaiheen toiminnan arvioimisesta epävarmaa.

Vaikka 1. vaiheen herkkyys oleellisten lausuntojen havaitsemisessa oli tämän työn puitteissa hyvä, virketasolla herkkyys oli alhainen. Herkkyyttä voitaisiin parantaa edelleen laajentamalla käytettyä sanalista.

Lausuntoaineisto koostui viidestä eri kuvantamismodaliteetista: TT, PET-TT, SPET-TT, MRI ja gammakuvaus. Käytetyllä lausuntoaineistolla näistä muodostui yhteensä 33 erilaista kuvantamistutkimusta, joista huomattava osa oli syövän levinneisyyden kannalta joko epäolennaisia tai erikoistutkimuksia tavanomaisten levinneisyystutkimuksien jälkeen. Lisäksi työssä käytetty lausuntoaineisto koostui ainoastaan melanoomapotilaista, vaikka malli ei ole suunnattu vain yhden syöpätyypin tunnistamista varten. Laadultaan ja määrältään rajoittunut aineisto sekä kuvantamistutkimuksien heterogeenisyys tekevät tuloksista vaikeasti yleistettäviä. Kuitenkin 1. vaiheen toimiminen epätyypillisellä aineistolla on myönteinen tulos, sillä mallin on tarkoitus olla kuvantamismodaliteetista ja syöpätyypistä riippumaton. Mallin 1. vaiheen testaaminen suuremmalla aineistolla, joka koostuisi useista eri syöpätyypeistä, antaisi luotettavamman arvion sen kyvystä tunnistaa oleellisia lausuntoja. Lisäksi 1. vaiheen testaaminen aineistolla, joka koostuisi vain levinneisyystutkimuksista, mittaisi paremmin mallin toimintaa sen suunniteltua käyttökohdetta ajatellen.

4.1 Johtopäätökset

Tässä työssä validoimme Auria Tietopalveluiden kehittämän mallin 1. vaiheen kykyä tunnistaa metastaasitietoa radiologien lausunnoista. Käytetyllä lausuntoaineistolla mallin 1. vaihe kykenee poimimaan oleellisen tiedon lausuntotasolla hyvin herkästi, mutta 1. vaiheen toiminta vaatii laajempaa testaamista ja validointia isommalla aineistolla sekä todennäköisesti kattavampaa sanalista, jotta herkkyys havaita oleellisia lausuntoja paranisi.

Lähteet

- Nobel JM, Puts S, Bakers FCH, Robben SGF, Dekker ALAJ. “Natural Language Processing in Dutch Free Text Radiology Reports: Challenges in a Small Language Area Staging Pulmonary Oncology.” *Journal of Digital Imaging*. 2020;33(4):1002-1008
- Babita Pandey, Devendra Kumar Pandey, Brijendra Pratap Mishra, Wasiur Rhmann “A comprehensive survey of deep learning in the field of medical imaging and medical natural language processing: Challenges and research directions” *Journal of King Saud University - Computer and Information Sciences*, Volume 34, Issue 8, Part A, 2022, Pages 5083-5099
- Pons E, Braun LM, Hunink MG, Kors JA. “Natural Language Processing in Radiology: A Systematic Review” *Radiology* vol. 279,2 (2016): 329-43
- Casey A, Davidson E, Poon M, Dong H, Duma D, Grivas A, Grover C, Suárez-Paniagua V, Tobin R, Whiteley W, Wu H, Alex B. “A systematic review of natural language processing applied to radiology reports” *BMC Medical Informatics and Decision Making* 21, 179 (2021)
- Ahumada R, Dunstan J, Rojas M, Peñafiel S, Paredes I, Báez P. “Automatic Detection of Distant Metastasis Mentions in Radiology Reports in Spanish” *JCO Clinical Cancer Informatics*. vol. 8 e2300130 (2024)
- Soysal E, Warner JL, Denny JC, Xu H. “Identifying Metastases-related Information from Pathology Reports of Lung Cancer Patients.” *AMIA Joint Summits on Translational Science Proceedings* vol. 2017 268-277. 26 Jul. 2017
- Banerjee I, Bozkurt S, Caswell-Jin JL, Kurian AW, Rubin DL. “Natural Language Processing Approaches to Detect the Timeline of Metastatic Recurrence of Breast Cancer” *JCO Clinical Cancer Informatics*. 2019 Oct;3:1-12
- Karimi YH, Blayney DW, Kurian AW, Shen J, Yamashita R, Rubin D, Banerjee I. “Development and Use of Natural Language Processing for Identification of Distant Cancer Recurrence and Sites of Distant Recurrence Using Unstructured Electronic Health Record Data” *JCO Clinical Cancer Informatics*. 2021 Apr;5:469-478
- Groot, O. Q., Bongers, M. E. R., Karhade, A. V., Kapoor, N. D., Fenn, B. P., Kim, J., Verlaan, J. J., & Schwab, J. H. (2020). “Natural language processing for automated quantification of bone metastases reported in free-text bone scintigraphy reports” *Acta oncologica* (Stockholm, Sweden), 59(12), 1455–1460

- Senders JT, Karhade AV, Cote DJ, et al. “Natural Language Processing for Automated Quantification of Brain Metastases Reported in Free-Text Radiology Reports” *JCO Clinical Cancer Informatics*. 2019;3:1-9
- Yang R, Zhu D, Howard LE, et al. “Identification of Patients With Metastatic Prostate Cancer With Natural Language Processing and Machine Learning” *JCO Clinical Cancer Informatics*. 2022;6:e2100071
- Park HJ, Park N, Lee JH, et al. “Automated extraction of information of lung cancer staging from unstructured reports of PET-CT interpretation: natural language processing with deep-learning” *BMC Medical Informatics and Decision Making*. 2022;22(1):229.
- Pan Q, Zhao F, Chen X, Chen D. “A method for extracting tumor events from clinical CT examination reports” *Journal of Biomedical Informatics*. 2023;142:104371