Radiomics-based machine learning classification of bone chondrosarcoma

Gitto, S.

Samenvatting en algemene discussie

Samenvatting

Hoofdstuk 1 is een algemene inleiding tot het doctoraal onderzoek. Het doel van dit proefschrift was het bepalen van de diagnostische waarde van kunstmatige intelligentie (in deze Nederlandstalige samenvatting wordt de meer algemene term kunstmatige intelligentie gebruikt om de specifiekere term machine learning te vertalen) bij het differentiëren tussen atypische cartilagineuze tumor (ACT) en hooggradig chondrosarcoom (CS), gebaseerd op radiomics kenmerken die zijn afgeleid van dwarsdoorsnedebeeldvorming, zoals magnetische resonantie beeldvorming (MRI) en computertomografie (CT), in vergelijking met ervaren radiologen gespecialiseerd in musculoskeletale oncologie.

In hoofdstuk 2 introduceerden we het concept van CT- en MRI-radiomics van bot- en weke delen sarcomen door middel van een literatuurstudie van de reproduceerbaarheid van radiomics karakteristieken en validatiestrategieën voor voorspellende modellen. Het uiteindelijke doel van deze systematische evaluatie was om een consensus te bereiken over deze aspecten die van belang zijn voor de translatie en toepasbaarheid van radiomics in de klinische praktijk. Van de 278 geïdentificeerde artikelen werden negenenveertig artikelen opgenomen die tussen 2008 en 2020 werden gepubliceerd. Ze behandelden radiomics van bottumoren (n=12) of tumoren van de weke delen (n=37). Achttien (37%) onderzoeken bevatten een analyse van de reproduceerbaarheid van de berekende radiomics karakteristieken. De variabiliteit van de beeldsegmentatie gedaan door dezelfde persoon (intra-observervariabiliteit) en tussen verschillende personen (interobserver variabiliteit) was het thema van de reproduceerbaarheidsanalyse in 16 (33%) onderzoeken, hetgeen meer was dan het aantal analyses die gericht waren op beeldacquisitie of beeldbewerking (n=2, 4%). De intraclass correlatiecoëfficiënt (ICC) was de meest gebruikte statistische methode om de reproduceerbaarheid te beoordelen. Deze varieerde van 0.6 tot 0.9. In 25 (51%) artikelen werd ten minste één validatietechniek voor kunstmatige intelligentie gebruikt ten behoeve van modelontwikkeling. K-voudige kruisvalidatie werd het meest gebruikt. Een klinische validatie van het model werd gerapporteerd in 19 (39%) artikelen. Dit werd uitgevoerd met behulp van een separate dataset afkomstig van de primaire instelling (d.w.z. interne validatie)

in 14 (29%) studies en met een onafhankelijke dataset waarbij data werden gegenereerd met verschillende scanners of in een andere instelling (d.w.z externe validatie) in 5 (10%) studies. Concluderend bleek dat de reproduceerbaarheid van radiomics kenmerken en modelvalidatie tussen de studies over musculoskeletale sarcomen in hoge mate varieerden. Dit zou in toekomstige onderzoeken moeten worden aangepakt om het gebied van radiomics van een onderzoeksgebied naar het klinische domein te brengen.

In hoofdstuk 3 werd de diagnostische waarde van op MRI-radiomics gebaseerde modellering met kunstmatige intelligentie (specifiek machine learning) gevalueerd in het onderscheiden van ACT en hooggradig CS in een eerste studie in één centrum. We hebben retrospectief 58 patiënten bestudeerd met een histologisch bewezen ACT (n=26) of een hooggradig CS (n=32, inclusief 16 gelokaliseerd in het appendiculaire en 16 in het axiale skelet). Ze werden willekeurig verdeeld in trainings- (n=42) en test- (n=16) groepen voor respectievelijk de ontwikkeling en validatie van het model. Alle tumoren werden handmatig gesegmenteerd op T1- en T2-gewogen MRI door 2D interessegebieden (Regions Of Interest: ROIs) te tekenen die werden gebruikt voor extractie van radiomics kenmerken. Na het selecteren van een functie werd een ensemble-classificeerder (AdaBoostM1) op de training set afgestemd met behulp van 10-voudige kruisvalidatie en vervolgens werd getest op een niet eerder gebruikte testset. Daarna heeft een ervaren musculoskeletale oncologische radioloog geblindeerd voor histologie en radiologische gegevens de laesies in de testgroep kwalitatief geëvalueerd. De gegevensset werd na functieselectie teruggebracht tot 4 T1-gewogen MRI-radiomics karakteristieken. Het model identificeerde respectievelijk 85.7% (AUC=0.85) en 75% (AUC=0.78) van de laesies in de trainings- en testgroepen. De radioloog identificeerde 81.3% van de laesies correct, zodat er geen verschil was met het model (p=0.453). Concluderend toonde dat ons ontwikkelde model gebaseerd op kunstmatige intelligentie accuraat het onderscheid tussen ACT en hooggradig CS kon maken.

In hoofdstuk 4 werd de invloed onderzocht van de handmatige segmentatievariabiliteit tussen observatoren op de reproduceerbaarheid van 2D en 3D CT- en MRI-gebaseerde textuuranalyse. Dertig patiënten met kraakbeentumoren (n=10 enchondroom; N=10 ACT; N=10 hooggradig CS) werden retrospectief opgenomen. Drie radiologen voerden onafhankelijk van elkaar handmatige beeldsegmentatie uit op natieve CT, T1-gewogen en T2-gewogen MRI door zowel een 2D ROI op de coupe te tekenen die het grootste tumoroppervlak toonde als een 3D ROI te tekenen die het gehele tumorvolume

omlijnde. Bovendien werd een marginale reductie van de ROI toegepast op zowel 2D- als 3D-segmentaties om de invloed van segmentatiemarges te evalueren. In totaal werden 783 en 1132 radiomics kenmerken geëxtraheerd uit respectievelijk originele en gefilterde 2D- en 3D-beelden. Functiestabiliteit werd gedefinieerd als een ICC≥0.75. In 2D versus 3D beeldsegmentatie waren de waarden van stabiele eigenschappen voor respectievelijk CT-, T1-gewogen en T2-gewogen beelden 74.71% versus 86.57% ($p<0.001$), 77.14% versus 80.04% ($p=0.142$) en 95.66% versus 94.97% ($p=0.554$). De reductie van de marge verbeterde 2D segmentatie ($p=0.343$) niet en presteerde slechter dan 3D ($p<0.001$) beeldsegmentatie wat betreft stabiliteit van de radiomics kenmerken. Bij 2D- versus 3D-beeldsegmentatie bedroegen de overeenkomende stabiele kenmerken die waren afgeleid van CT en MRI 65.8% versus 68.7% ($p=0.191$), en die afgeleid van T1-gewogen en T2-gewogen beelden 76.0% versus 78.2% ($p=0.285$). Concluderend waren 2D en 3D CT en MRI radiomics kenmerken van kraakbeentumoren reproduceerbaar, hoewel een zekere mate van segmentatievariabiliteit tussen observatoren de noodzaak van betrouwbaarheidsanalyse in radiomics studies aantoonde.

In hoofdstuk 5 werd een multicentrische studie beschreven die de prestaties van CT op radiomics-gebaseerde kunstmatige intelligentie modellen onderzocht in het onderscheiden tussen ACT en hooggradig CS van lange pijpbeenderen. Honderdtwintig patiënten met histologie-bewezen laesies werden retrospectief opgenomen. Het trainingscohort bestond uit 84 CT-scans uit centrum 1 (n=55 ACT; N=29 CS graad II-IV). Het externe testcohort bestond uit de CT-beelden afkomstig van 36 PET-CT-scans uit centrum 2 (n=16 ACT; N=20 CS graad II-IV). 2D-segmentatie werd uitgevoerd op preoperatieve CT-scans. Radiomics kenmerken werden geëxtraheerd. Na reductie van datadimensionaliteit en uitvoer van een klassenbalans in centrum 1 werd een kunstmatige intelligentie gebaseerde classificeerder (LogitBoost) gevalideerd van een intern testcohort (tot stand gekomen met behulp van 10-voudige kruisvalidatie) en op het externe testcohort. In centrum 2 werden, m.b.v. de McNemar test, de resultaten vergeleken met die van een preoperatieve biopsie en met de beoordeling van een ervaren radioloog. Het model had een nauwkeurigheid van 81% (AUC=0.89) en 75% (AUC=0.78) bij het identificeren van de laesies in respectievelijk de training en de externe testcohorten. In het bijzonder was de nauwkeurigheid bij het classificeren van ACT en hooggradig CS respectievelijk 84% en 78% in het trainingscohort, en 81% en 70% in het externe testcohort. Preoperatieve biopsie had

een nauwkeurigheid van 64% (AUC=0.66) (p=0.29). De radioloog bereikte een nauwkeurigheid van 81% (p=0.75). Conclusie was dat het model een goede nauwkeurigheid bereikte bij het classificeren van ACT en hooggradige CS van lange pijpbeenderen op basis van radiomics kenmerken afkomstig van preoperatieve CT-scans.

In hoofdstuk 6 werd een multicentrische studie beschreven die de prestaties van MR op radiomics-gebaseerde kunstmatige intelligentie onderzocht in het onderscheiden tussen ACT en hooggradig CS van lange pijpbeenderen. Honderdachtenvijftig patiënten uit twee tertiaire centra voor bottumoren met chirurgisch behandelde en histologisch bewezen kraakbeentumoren werden retrospectief geïncludeerd. Het trainingscohort bestond uit 93 MRI-scans uit centrum 1 (n=74 ACT; N=19 CS graad II). Het externe testcohort bestond uit 65 MRI-scans uit centrum 2 (n=45 ACT; N=20 CS graad II). 2D-segmentatie werd handmatig uitgevoerd op T1-gewogen MRI-sequenties. Eerste orde, morfologische- en textuureigenschappen werden geëxtraheerd. Reductie van datadimensionaliteit werd uitgevoerd op basis van stabiliteit, variatie en inter-correlatie analyses en recursieve kenmerk eliminatie op de data afkomstig van centrum 1 na balancering van klassen (CS graad II oversampling naar N=74). Zo werd een model gebaseerd op kunstmatige intelligentie (Extra Trees Classifier) getraind op een cohort tot stand gekomen door 10-voudige kruis-validatie en getest op een extern test cohort. In centrum 2 werden de prestaties, m.b.v. de McNemar's test, vergeleken met die van een radioloog ervaren op het gebied van musculoskeletale oncologie. Negenhonderdnegentien radiomics kenmerken werden geëxtraheerd en vervolgens teruggebracht tot 17 door middel van datadimensionaliteitsvermindering. Na training had het model (AUC=0.88), een nauwkeurigheid van 92% (60/65, AUC=0.94) voor het identificeren van de laesies in het externe testcohort. In het bijzonder waren de nauwkeurigheden voor het correct classificeren van ACT en graad II CS respectievelijk 98% (44/45) en 80% (16/20). De ervaren radioloog had een nauwkeurigheid van 98% (64/65) en er was geen significant verschil met het getrainde model (p>0.99). Concluderend bereikte het model een hoge nauwkeurigheid voor het classificeren van ACT en graad II CS van lange pijpbeenderen op basis van MRI-radiomics kenmerken.

Algemene discussie, beperkingen en toekomstperspectieven

Dit proefschrift richt zich op het concept ACT, dat is gedefinieerd volgens de 2013 en 2020 classificatie van de Wereldgezondheidsorganisatie [1,2]. Deze relatief nieuwe

definitie weerspiegelt het indolente biologische gedrag van ACT, dat nu wordt beschouwd als een intermediaire kraakbeentumor van lange pijpbeenderen in plaats van een maligne tumor [2]. Dit concept sluit beter aan bij therapeutische opties die voor ACT volledig verschillen met die van hooggradig (II of hoger) appendiculair CS en axiaal CS van elke graad [3]. Wijde resectie met tumorvrije grenzen is de therapie van keuze voor de laatste groep. De behandeling van ACT is de afgelopen drie decennia opmerkelijk veranderd. Tot in de jaren negentig werd er een wijde resectie uitgevoerd, waarna de therapie veranderde in intra-lesionale curettage en tegenwoordig is in toenemende mate waakzaam afwachten met beeldvorming, maar zonder chirurgisch ingrijpen een optie [4].

Vanwege het risico van het niet verkrijgen van representatief biopsiemateriaal afkomstig uit het meest kwaadaardige deel van de tumor, wordt een biopsie in veel tertiaire centra niet langer gebruikt [5]. Gezien de toenemende incidentie van ACT als gevolg van een toename van incidentele bevindingen op MRI [6] is er behoefte aan duidelijke beeldvormingscriteria om onderscheid te maken tussen ACT enerzijds en hooggradige CS anderzijds. Beoordeling van beeldvorming lijdt echter aan interobserver-variabiliteit [7,8]. In dit proefschrift worden modellen op basis van radiomics kenmerken voorgesteld om op basis van beeldvorming het onderscheid tussen ACT en hooggradig CS objectiever te maken. Andere radiomics studies tot nu toe hebben zich gericht op het differentiëren tussen enchondroom enerzijds en ACT/hooggradig CS anderzijds [9–11]. De differentiatie tussen enchondroom en ACT is echter geleidelijk minder relevant geworden, vanwege de hierboven genoemde herwaardering van ACT en de daarbij horende nieuwe inzichten omtrent behandeling waarbij het controleren van de laesie met behulp van beeldvorming een alternatief is voor het curetteren van de laesie [4,12,13]. Resectie van ACT is, in tegenstelling tot behandeling van CS graad II, niet meer geïndiceerd. Aangezien enchondroom en hooggradig CS eenvoudig te onderscheiden zijn met behulp van röntgenfoto's, heeft het diagnostisch dilemma zich verplaatst van onderscheid tussen enchondroom en ACT naar onderscheid tussen ACT en CS graad II [4].

In de hoofdstukken 3, 5 en 6 van dit proefschrift werden kunstmatige intelligentie gebaseerde technieken gebruikt om classificatiemodellen te maken op basis van radiomics kenmerken die zijn afgeleid van MRI en CT. In de hoofdstukken 5 en 6 werden met name grote multicentrische studies beschreven, inclusief klinische validatie van het model op onafhankelijke data (externe testcohort) van een andere instelling. Deze studies

concentreerden zich op respectievelijk CT- en MRI-radiomics en behaalden een hoge nauwkeurigheid bij het correct classificeren van ACT en hooggradig CS van lange pijpbeenderen, zonder verschil met ervaren radiologen op het gebied van musculoskeletale oncologie. Ongetwijfeld werd de meest relevante bevinding gedaan in hoofdstuk 6, waarbij een getraind model op basis van radiomics kenmerken afkomstig uit T1-gewogen MRI een nauwkeurigheid van 92% bereikte bij het onderscheiden tussen ACT en CS graad II. Bovendien was er geen statistisch significant verschil ($p>0.99$) tussen resultaten van het model en die van een radioloog, gespecialiseerd in bottumoren met 35 jaar ervaring. In hoofdstuk 4 hebben we de invloed van de segmentatievariabiliteit tussen observatoren op de reproduceerbaarheid van 2D en 3D CT- en MRI-radiomics kenmerken van kraakbeentumoren methodologisch geanalyseerd. Er was een goede reproduceerbaarheid van radiomics kenmerken op alle beeldvormende modaliteiten, zowel van kenmerken op basis van 2D- als 3D-segmentaties, hoewel voorafgaande kennis m.b.t. stabiliteit van radiomics kenmerken een belangrijke voorwaarde is om nauwkeurige classificatie te bewerkstelligen. Dit werd ook benadrukt in de systematische review van hoofdstuk 2 m.b.t. reproduceerbaarheid van kenmerken en validatie strategieën, en dit werd ook gedaan in onze daaropvolgende studies (hoofdstukken 5 en 6).

Dit proefschrift kent enkele beperkingen. In de eerste plaats waren alle uitgevoerde onderzoeken retrospectief, omdat hiermee een groter aantal patiënten (met een relatief zeldzame aandoening) kon worden geïncludeerd. Bovendien is een prospectieve analyse niet strikt nodig in radiomics studies [14]. In de tweede plaats kan de segmentatietechniek de resultaten beïnvloeden. In onderzoeken die in hoofdstukken 3, 5 en 6 worden beschreven hebben we een 2D-segmentatie uitgevoerd op de coupe geselecteerd met de grootste tumordimensie. Een 2D-benadering had de voorkeur, omdat het potentieel gemakkelijker is om dit in de klinische praktijk te implementeren, bovendien werd in recente literatuur gemeld dat 2D-segmentatie betere resultaten zou opleveren dan 3D-segmentatie [15]. Onze bevindingen in hoofdstuk 4 toonden geen verschil tussen 2D en 3D MRI-gebaseerde textuuranalyse. Ten derde was ACT oververtegenwoordigd in vergelijking met hooggradig CS in hoofdstuk 5, vooral in het trainingscohort uit centrum 1, en in hoofdstuk 6. Dit weerspiegelde echter nauwkeurig de incidentie van ACT en hooggradig CS [6], bovendien werd balansering uitgevoerd in beide studies om de minderheidsklasse kunstmatig te oversampelen in de trainingscohorten [16]. Ten vierde werden geen contrastmiddelen

gebruikt bij CT en MRI voor de ontwikkelde modellen, omdat deze niet bij alle patiënten beschikbaar waren. Tot slot is het gebruik van de lage dosis CT-beelden van PET-CT in het externe testcohort zoals beschreven in hoofdstuk 5 suboptimaal. De goede resultaten stemmen ons echter positief en openen de mogelijkheid voor toekomstig onderzoek om licht te werpen op de waarde van het gebruik van kunstmatige intelligentie modellen die gebruik maken van radiomics kenmerken afkomstig van CT- of MRI.

Concluderend toonden CT- en MRI-modellen die gebruik maken van radiomics kenmerken een hoge nauwkeurigheid aan bij het onderscheiden van ACT en hooggradig CS en ziet het er veelbelovend uit als een objectieve beeldvormingsmethode die kan worden gebruikt bij de klinische besluitvorming. Dit kan met name belangrijk zijn in de algemene praktijk waar gespecialiseerde expertise niet voorhanden is, bij het correct identificeren van de veel voorkomende ACT. Onze grote onderzoekspopulatie en de zeer goede prestaties die zijn bereikt met onafhankelijke gegevens van verschillende instellingen, zoals gepresenteerd in de hoofdstukken 5 en 6, garanderen de generaliseerbaarheid van onze resultaten. Toekomstige studies zullen de toepasbaarheid van onze bevindingen in de klinische praktijk moeten verifiëren.

Referenties

[1] Fletcher CDM, Bridge JA, Hogendoorn PCW, Mertens F. World Health Organization Classification of Tumours of Soft Tissue and Bone. Lyon, France: International Agency for Research on Cancer Press; 2013.

[2] WHO Classification of Tumours Editorial Board. WHO Classification of Tumours: Soft Tissue and Bone Tumours. Lyon, France: International Agency for Research on Cancer Press; 2020.

[3] Casali PG, Bielack S, Abecassis N, Aro HT, Bauer S, Biagini R, et al. Bone sarcomas: ESMO–PaedCan–EURACAN Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol 2018;29:iv79–95.

[4] van de Sande MAJ, van der Wal RJP, Navas Cañete A, van Rijswijk CSP, Kroon HM, Dijkstra PDS, et al. Radiologic differentiation of enchondromas, atypical cartilaginous tumors, and high-grade chondrosarcomas—Improving tumor-specific treatment: A paradigm in transit? Cancer 2019;125:3288–91.

[5] Hodel S, Laux C, Farei-Campagna J, Götschi T, Bode-Lesniewska B, Müller DA. The impact of biopsy sampling errors and the quality of surgical margins on local recurrence and survival in chondrosarcoma. Cancer Manag Res 2018;10:3765–71.

[6] van Praag (Veroniek) VM, Rueten-Budde AJ, Ho V, Dijkstra PDS, Fiocco M, van de Sande MAJ, et al. Incidence, outcomes and prognostic factors during 25 years of treatment of chondrosarcomas. Surg Oncol 2018;27:402–8.

[7] Jones KB, Buckwalter JA, McCarthy EF, DeYoung BR, El-Khoury GY, Dolan L, et al. Reliability of Histopathologic and Radiologic Grading of Cartilaginous Neoplasms in Long Bones. J Bone Joint Surg Am 2007;89:2113–23.

[8] Zamora T, Urrutia J, Schweitzer D, Amenabar PP, Botello E. Do Orthopaedic Oncologists Agree on the Diagnosis and Treatment of Cartilage Tumors of the Appendicular Skeleton? Clin Orthop Relat Res 2017;475:2176–86.

[9] Fritz B, Müller DA, Sutter R, Wurnig MC, Wagner MW, Pfirrmann CWA, et al. Magnetic Resonance Imaging–Based Grading of Cartilaginous Bone Tumors. Invest Radiol 2018;53:663–72.

[10] Lisson CS, Lisson CG, Flosdorf K, Mayer-Steinacker R, Schultheiss M, von Baer A, et al. Diagnostic value of MRI-based 3D texture analysis for tissue characterisation and discrimination of low-grade chondrosarcoma from

enchondroma: a pilot study. Eur Radiol 2018;28:468–77.

[11] Pan J, Zhang K, Le H, Jiang Y, Li W, Geng Y, et al. Radiomics Nomograms Based on Non-enhanced MRI and Clinical Risk Factors for the Differentiation of Chondrosarcoma from Enchondroma. J Magn Reson Imaging 2021; 54:1314–23.

[12] Deckers C, Schreuder BHW, Hannink G, de Rooy JWJ, van der Geest ICM. Radiologic follow-up of untreated enchondroma and atypical cartilaginous tumors in the long bones. J Surg Oncol 2016;114:987–91.

[13] Omlor GW, Lohnherr V, Lange J, Gantz S, Mechtersheimer G, Merle C, et al. Outcome of conservative and surgical treatment of enchondromas and atypical cartilaginous tumors of the long bones: retrospective analysis of 228 patients. BMC Musculoskelet Disord 2019;20:134.

[14] Lubner MG, Smith AD, Sandrasegaran K, Sahani D V., Pickhardt PJ. CT Texture Analysis: Definitions, Applications, Biologic Correlates, and Challenges. Radiographics 2017;37:1483–503.

[15] Ren J, Yuan Y, Qi M, Tao X. Machine learning–based CT texture analysis to predict HPV status in oropharyngeal squamous cell carcinoma: comparison of 2D and 3D segmentation. Eur Radiol 2020;30:6858–66.

[16] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res 2002;16:321–57.