Automated quality assurance of deep learning contours in head-and-neck radiotherapy

Mody, P.P.

7

Samenvatting, discussie en toekomstig werk

7.1 Samenvatting van de dissertatie

Deze dissertatie richt zich op de dringende behoefte aan efficiënte en betrouwbare kwaliteitscontrole (QA)-tools voor geautomatiseerde contourbepaling van organen en tumoren bij radiotherapie. Hoewel deep learning modellen een aanzienlijke versnelling bieden, kunnen de daaropvolgende handmatige QA- en verfijningsstappen tijdrovend zijn en zo de winst gedeeltelijk tenietdoen, wat leidt tot een knelpunt in klinische workflows. Twee hoofdthema's binnen QA worden onderzocht: *foutdetectie* (waar zijn contouren waarschijnlijk fout) en *foutcorrectie* (hoe corrigeer je die efficiënt), in zowel pre- als post-commissioning fasen.

Dit proefschrift onderzoekt specifiek: a) de ontwikkeling van een geautomatiseerde en schaalbare workflow voor het evalueren van de dosimetrische impact van autocontouren vóór ingebruikname (Chapter 2), b) het potentieel van Bayesiaanse modellen en trainingsverliezen om onnauwkeurige voorspellingen te detecteren in de post-ingebruiknamefase door gebruik te maken van de bijbehorende onzekerheid (Chapter 3 & Chapter 4), en c) de verbetering van de efficiëntie van foutcorrectie met behulp van AI-ondersteunde verfijningstools (Chapter 5). Het overkoepelende doel van dit proefschrift is dan ook om verschillende QA-methodologieën te verkennen, zowel vóór als na ingebruikname van autocontouringtools voor hoofd-halsradiotherapie.

7.2 Hoofdstuk Samenvattingen

7.2.1 Hoofdstuk 2

Dit hoofdstuk richtte zich op de noodzaak van grootschalige pre-commissioning dosimetrische evaluaties van automatisch gecontourde organen-at-risk (OARs). Het belangrijkste resultaat was een geautomatiseerde workflow voor planningsoptimalisatie, gebaseerd op bestaande klinische instellingen, bekend als robotic process automation (RPA). Via scripting in het Treatment Planning System (TPS) werd een handmatige planningsprocedure geautomatiseerd.

Een studie werd uitgevoerd op 100 hoofd-hals patiënten (70 fotonen, 30 protonen). Resultaten toonden minimale dosimetrische verschillen tussen automatische en handmatige contouren. Dit wijst erop dat geometrische afwijkingen veroorzaakt door automa-

tische contouren beperkte klinische impact hebben, en bevestigt de bruikbaarheid van de voorgestelde QA-aanpak.

7.2.2 Hoofdstuk 3

Dit hoofdstuk onderzocht hoe modelkeuzes de onzekerheid beïnvloeden, die als proxy kan dienen voor fouten in post-commissioning QA. Twee Bayesiaanse modellen (DropOut en FlipOut) werden geëvalueerd met behulp van Expected Calibration Error (ECE) en een nieuwe metriek, Region-based Accuracy-vs-Uncertainty (R-AvU). Waar ECE een informatietheoretische benadering is, biedt R-AvU een meer visuele evaluatie. Training met cross-entropy verlies (CE) gaf betere calibratie (lagere ECE). FlipOut-CE toonde betere onzekerheidsdekking in foutieve regio's dan DropOut-CE volgens de R-AvU grafieken. Deze resultaten roepen de vraag op: welke metriek moet men gebruiken bij onzekerheidsevaluatie voor foutdetectie?

7.2.3 Hoofdstuk 4

Hoewel Bayesiaanse modellen onzekerheidskaarten kunnen produceren, is hun klinisch nut afhankelijk van de mate waarin deze kaarten overeenkomen met echte fouten. Hoofdstuk 2 toonde aan dat deze overeenstemming vaak suboptimaal is. Dit hoofdstuk introduceerde een differentieerbaar verlies op basis van de Accuracy-vs-Uncertainty (AvU) metriek, die expliciet onzekerheid stimuleert waar fouten voorkomen. De kaarten werden geëvalueerd via ROC-curves ("uncertainty-ROC") en Precision-Recall-curves. Een belangrijk aspect was het onderscheid tussen "fouten" (klein, acceptabel) en "falen" (groter, vereisen interventie).

De AvU-verliesfunctie verbeterde significante calibratie (ECE) en de overeenkomst tussen onzekerheid en fouten (ROC-AUC, PRC-AUC), voor zowel in-distributie (ID) als out-of-distributie (OOD) datasets. AvU presteerde zelfs beter dan ensemble-modellen. Dit toont dat optimalisatie op ECE niet voldoende is om bruikbare onzekerheidskaarten te produceren — AvU biedt een unieke meerwaarde.

7.2.4 Hoofdstuk 5

De focus ligt hier op foutcorrectie, post-commissioning. In dit hoofdstuk werd de efficiëntie en kwaliteit van handmatige borstels vergeleken met een AI-ondersteunde "AI potlood"-tool. Bestaande AI-potloodtools missen vaak evaluatie met menselijke gebruikers en werken enkel in 2D.

Een webinterface werd ontwikkeld waarin gebruikers 2D-aanduidingen (scribbles) konden geven, waarna de AI potlood 3D-refinements uitvoerde op CT+PET scans van hoofd-hals tumoren. Zowel klinische als niet-klinische gebruikers namen deel. De AI-potlood was 5–78% sneller bij niet-experts en 16–97% sneller bij experts, terwijl de uiteindelijke kwaliteit gelijkwaardig bleef. De AI potlood bereikte snel een hoge kwaliteit, in tegen-

stelling tot de geleidelijke verbetering bij de handmatige tool. Dit toont de kracht van AI-geassisteerde QA bij radiotherapie.

7.3 Discussie en toekomstig werk

Deze dissertatie adresseert belangrijke uitdagingen in veilige, efficiënte en betrouwbare integratie van QA-tools voor deep learning-gebaseerde auto-contouring in de kliniek. Zowel foutdetectie als foutcorrectie, in pre- en post-commissioning scenario's, worden aangepakt, met nadruk op menselijke bruikbaarheid.

Toekomstige onderzoekslijnen zijn:

- Klinische betrokkenheid - Technisch onderzoek probeert vaak te optimaliseren op basis van bepaalde, vooraf gespecificeerde criteria, maar vertaalt dit niet naar de klinische praktijk. Dit gebrek aan 'van tafel tot bed'-mentaliteit wordt vaak veroorzaakt door de structuur van onderzoeksprojecten. Een ontbrekende factor is vaak voldoende klinische betrokkenheid, waardoor onderzoek op een stoffige plank blijft liggen. Onderzoekers zouden moeten overwegen hun teams en mentoren zo in te richten dat ze multidisciplinaire vaardigheden inzetten om de volledige breedte en diepte van het probleem te begrijpen.

- Vernieuwing van contourrichtlijnen – Chapter 2 toonde zowel correlaties als non-correlaties tussen DICE en dosisverschillen. Grotere studies zouden de contourrichtlijnen opnieuw kunnen definiëren, en mogelijk vaste anatomische richtlijnen laten evolueren naar richtlijnen met marges die rekening houden met inter- en intra-observatorvariabiliteit.

- Het nut van onzekerheid in klinische settings begrijpen – Onzekerheid is een wiskundig concept dat de potentie heeft om inzicht te bieden in de betrouwbaarheid van datagestuurde technieken zoals deep learning. De community gebruikt echter vaak puur wiskundige concepten zoals ECE (met zijn groeperingsmechanisme) om het nut van de onzekerheid van een model te evalueren. Dergelijke statistieken leggen onzekerheid niet pixelgewijs (of op een gedetailleerde manier) vast. Het verleggen van de grenzen van bestaande statistieken, hoe belangrijk ook, is dus niet voldoende om onderzoeksinnovaties aan te passen aan de dagelijkse klinische praktijk.

- Pixel- vs. slice- vs. regio-onzekerheid: Het is mogelijk dat er een praktische limiet is aan de mate waarin clinici kunnen profiteren van 'onzekerheidsafstemming' voordat het leidt tot cognitieve overbelasting. Enerzijds kan te veel op onzekerheid gebaseerde besluitvorming (bijv. per pixel) cognitief veeleisend zijn. Anderzijds biedt gemiddelde onzekerheid (bijv. op het niveau van een plak of orgaan/tumor) mogelijk niet effectief houvast voor het verfijnen van contouren. Daarom moeten

onderzoekers nadenken over de granulariteit van onzekerheid die we nodig hebben in medische beeldsegmentatietoepassingen.

- Verliesfuncties en klinische bruikbaarheid: De DICE-loss is een geometrie-gebaseerde verliesfunctie, omdat het kijkt naar de algehele structuur en vorm van de 'ground truth' en de voorspelling. Verrassend genoeg presteerde een pixel-gebaseerde aanpak, namelijk de cross-entropy loss, echter beter in het weergeven van de werkelijke vertrouwelijkheid van zijn voorspellingen. Daarom moeten makers van autocontouring-tools dieper nadenken over de manier waarop hun verliesfuncties de ervaring van de eindgebruiker beïnvloeden.

- De eisen voor de dataset analyseren – Een van de belemmeringen voor het vertalen van onderzoek naar de klinische praktijk is de grote hoeveelheid benodigde trainingsdata. De literatuur laat echter vergelijkbare prestaties zien met datasets van verschillende groottes. Meer werk met hulpmiddelen zoals leercruves kan de gemeenschap beter informeren over de minimale vereisten voor datasets om te voldoen aan klinische normen voor het contouren van organen en doelwitten.

- Frameworks voor klinische validatie in de praktijk – Frameworks voor klinische validatie in de praktijk – Tools voor robuuste experimentatie en evaluatie zijn wat elk vakgebied vooruithelpt, aangezien ze de drempels voor nieuwkomers verlagen om bij te dragen. Dit is te zien bij programmeertalen zoals Python en deep learning frameworks zoals Tensorflow en PyTorch. Een vergelijkbaar voorbeeld voor medische beeldsegmentatie is het grand-challenge.org-platform. Nu deep learning tools steeds gangbaarder worden in de medische beeldvorming, moet de gemeenschap zich richten op de ontwikkeling van vergelijkbare frameworks voor onzekerheid als een proxy voor foutdetectie en voor interactieve segmentatie.

- Vertrouwen in AI-gedreven acties – Voor de context van interactieve contourverfijning, hoe kunnen we ervoor zorgen dat clinici de door AI gegenereerde verfijningen voldoende vertrouwen om niet terug te vallen op handmatige correcties? En kunnen dergelijke tools zich aanpassen aan de diverse manieren waarop verschillende clinici contourbewerking benaderen? Het kan dus nodig zijn om statistieken te gebruiken die bijhouden hoe betrouwbaar het model is in lokale gebieden waar de gebruiker zijn krabbels maakt. En doet het model stiekem onterechte voorspellingen in gebieden ver weg van de interactie van de gebruiker?.

- Rol van regelgevende instanties: Gezondheidszorgsystemen moeten worden gereguleerd door overheidsinstanties vanwege de kritieke aard van de dienst die ze leveren. Onderzoeks-innovaties overtreffen echter vaak de regulerende instanties en in de tussentijd bestaat de mogelijkheid dat innovaties die niet rigoureus of nauwkeurig

zijn getest, door artsen kunnen worden gebruikt. In het geval van op deep learning gebaseerde auto-contouring is er bijvoorbeeld heel weinig discussie over de noodzaak van land-/demografisch-gebaseerde benchmark datasets. Dit maakt het voor klinische innovators zeer omslachtig om te bepalen hoe ze commerciële oplossingen moeten evalueren, aangezien zij degene moeten zijn die hun eigen interne dataset samenstellen, die vaak rommelig is vanwege de drukke werkdruk van clinici. We smeken de lezer van dit proefschrift om over dit punt na te denken en de bovengenoemde lacune op te vullen..

7.4 Algemene conclusies

In een tijdperk van toenemende incidentie van kanker en beperkte klinische middelen, levert dit proefschrift essentiële hulpmiddelen om de veilige, effectieve integratie van deep learning auto-contouring in de radiotherapieworkflow te waarborgen. Door praktische, mensgerichte methoden te bieden voor zowel precieze foutdetectie als efficiënte foutcorrectie, helpt dit werk de kloof te overbruggen tussen geavanceerde deep learning-modellen en hun veilige en effectieve kwaliteitsbeoordeling voor integratie in de dagelijkse klinische radiotherapiepraktijk. We hopen anderen te inspireren om werk na te streven dat de kloof overbrugt tussen wiskundige onzekerheidsmetrieken en praktisch klinisch vertrouwen. Evenzo moeten interactieve AI-hulpmiddelen evolueren om de diverse manieren waarop clinici werken te weerspiegelen.

Uiteindelijk streeft dit onderzoek ernaar om patiëntenzorg van hoge kwaliteit te waarborgen en de workflowefficiëntie te verbeteren, waarbij de positieve resultaten bedoeld zijn om mensgerichte deep learning voor medische beeldvorming te informeren en te bevorderen.