

UNIVERSITEIT GENT

Faculteit Wetenschappen

---

# Deep learning voor autosegmentatie van computertomografie (CT) beelden in radiotherapie

---

MASTERPROEF VOORGELEGD VOOR HET BEHALEN VAN DE GRAAD VAN  
MASTER OF SCIENCE IN DE FYSICA EN DE STERRENKUNDE

*Auteur:*

De Rycke Jeffrey

*Promotor:*

Prof. Dr. ir. Barbara Vanderstraeten  
Vakgroep Structuur en Herstel van de Mens

*Copromotor:*

Prof. Dr. Luc Van Hoorebeke  
Vakgroep Fysica en Sterrenkunde

*Begeleiders:*

Dr. ir. Jan Aelterman  
Vakgroep Elektronica en Informatiesystemen

Dr. Eva Vandersmissen  
Agfa Radiology Solutions

Academiejaar: 2019-2020

## Samenvatting

Dit onderzoek heeft als doel het correct automatisch segmenteren van organen in CT-beelden aan de hand van een diep neurale netwerk voor gebruik binnen de radiotherapie. Het past enerzijds ideeën en *good practices* toe van vorige gelijkaardige onderzoeken zoals data-augmentatie, lossfuncties en metrieken, maar ook nieuwe ideeën zoals het beoordelen van de segmentatie door een klinische expert en het gebruik van klinische metrieken. We oordelen dat de finale resultaten competitief zijn met de theoretische limieten gezet door de variabiliteit in segmentaties van verschillende radiotherapeuten-oncologen en dat de segmentaties positief worden beoordeeld door een klinisch expert. Het model zou verder verfijnd kunnen worden door onze volledige dataset te gebruiken, te trainen op meer organen en/of in plaats van één CT-beeld, een kleine serie opeenvolgende CT-beelden te gebruiken om de segmentatie van het ene CT-beeld te voorspellen.

The goal of this research is to correctly automatically segment organs in CT-images for radiotherapy using a deep neural network. On the one hand, it uses ideas from previous similar studies such as data-augmentation, lossfunctions and metrics. On the other hand it uses new ideas as well, such as letting a clinical expert judge the segmentations and using clinical metrics. We conclude that the final results are competitive with the theoretical limits set by the variability in segmentations from different radiotherapists-oncologists and that the segmentations were positively received by a clinical expert. The model could be further refined by using our complete dataset, training on even more organs and/or instead of using only one CT-image, using a small series of consecutive CT-images to predict the segmentation of the one CT-image.



# Inhoudsopgave

<b>1</b>	<b>Inleiding</b>	<b>5</b>
1.1	Kadering van de thesis . . . . .	5
1.2	Radiotherapie . . . . .	6
1.3	Hounsfieldscalaal . . . . .	8
1.3.1	Definitie . . . . .	8
1.3.2	Radiodensiteit . . . . .	8
1.3.3	Attenuatiecoëfficiënt . . . . .	8
1.4	Deep learning in de medische wereld . . . . .	11
1.4.1	Een neuraal netwerk . . . . .	11
1.4.2	U-Net . . . . .	12
1.4.3	Gerelateerde onderzoeken . . . . .	13
1.5	Doel en probleemstelling . . . . .	15
1.5.1	Doel . . . . .	15
1.5.2	Organen en tumor . . . . .	15
<b>2</b>	<b>Methoden</b>	<b>17</b>
2.1	Datavoorbereiding . . . . .	17
2.1.1	Masker . . . . .	17
2.1.2	Datanormalisatie . . . . .	19
2.1.3	Data-augmentatie . . . . .	21
2.1.4	De HPC . . . . .	21
2.1.5	Geshuffelde schijnpatiënten . . . . .	22
2.2	De lossfunctie . . . . .	23
2.2.1	Diceloss . . . . .	23
2.2.2	Gewogen binaire crossentropie loss . . . . .	24
2.2.3	Gewogen categoriale crossentropie loss . . . . .	25
2.3	Pretrained encodergewichten van ImageNet . . . . .	26
2.4	Metriecken . . . . .	28
2.4.1	Dicescore . . . . .	28
2.4.2	IoU score . . . . .	29
2.4.3	DVH en clinical goals . . . . .	29
2.5	Trainen . . . . .	30
2.5.1	Gewogen binaire crossentropie loss . . . . .	30
2.5.2	Gewogen categoriale crossentropie loss . . . . .	31
2.5.3	Diceloss . . . . .	31
<b>3</b>	<b>Resultaten</b>	<b>32</b>
3.1	Pretrained ImageNet encoderweights . . . . .	32
3.2	Gewogen binaire crossentropie loss . . . . .	33
3.2.1	Longen . . . . .	33
3.2.2	Hart . . . . .	39
3.2.3	Slok darm . . . . .	43
3.2.4	Luchtpijp . . . . .	47
3.2.5	Ruggenmerg . . . . .	51
3.2.6	GTV . . . . .	55
3.3	Gewogen categoriale crossentropie loss . . . . .	63

3.3.1	Longen . . . . .	66
3.3.2	Hart . . . . .	68
3.3.3	Slokdarm . . . . .	70
3.3.4	Luchtpijp . . . . .	72
3.3.5	Ruggenmerg . . . . .	74
3.3.6	GTV . . . . .	76
3.4	Diceloss per orgaan . . . . .	78
3.4.1	Longen . . . . .	78
3.4.2	Hart . . . . .	82
3.4.3	Slokdarm . . . . .	86
3.4.4	Luchtpijp . . . . .	90
3.4.5	Ruggenmerg . . . . .	94
3.4.6	GTV . . . . .	98
3.5	Diceloss op alle organen . . . . .	102
3.5.1	Longen . . . . .	106
3.5.2	Hart . . . . .	108
3.5.3	Slokdarm . . . . .	110
3.5.4	Luchtpijp . . . . .	112
3.5.5	Ruggenmerg . . . . .	114
3.5.6	GTV . . . . .	116
3.6	DHV en clinical goals . . . . .	118
3.6.1	Visuele vergelijking . . . . .	119
3.6.2	Vergelijking van ROI (Region Of Interest) volumes. . . . .	123
3.6.3	Dosisverdeling . . . . .	126
3.6.4	DVH . . . . .	130
3.6.5	Clinical goals . . . . .	134
<b>4</b>	<b>Discussie</b>	<b>138</b>
4.1	Pretrained ImageNet encoderweights . . . . .	138
4.2	Gewogen binaire crossentropie loss . . . . .	138
4.3	Gewogen categoriale crossentropie loss . . . . .	139
4.4	Diceloss per orgaan . . . . .	140
4.5	Diceloss op alle organen . . . . .	141
4.6	DHV en clinical goals . . . . .	142
4.7	Vergelijking vorige onderzoeken . . . . .	143
4.8	Klinische metrieken . . . . .	143
4.9	Verder onderzoek . . . . .	143
<b>5</b>	<b>Conclusie</b>	<b>144</b>
<b>6</b>	<b>Dankwoord</b>	<b>145</b>
<b>A</b>	<b>Vergelijking originele en teststructuren in RayStation</b>	<b>148</b>

# 1 Inleiding

We zullen beginnen met het leveren van voldoende context en achtergrond bij deze thesis. Waarom kozen we voor dit onderwerp? Hoe ziet het huidige radiotherapeutische proces eruit? Welk werk en onderzoek is er al geleverd dat te maken heeft met ons doel? Wat willen wij concreet verwezenlijken en wat verwachten we? Volgende secties zullen deze vragen beantwoorden zodat de lezer voldoende achtergrondinformatie heeft over de huidige stand van zaken en wat wij willen onderzoeken.

## 1.1 Kadering van de thesis

Hoewel elke wetenschap hetzelfde doel heeft, namelijk het vergroten van ieders inzicht in hun respectievelijke onderzoeksvelden, leven ze vaak redelijk parallel aan elkaar. Men studeert vijf jaar en tracht de top van de top te bereiken in die richting. Echter valt er veel te halen uit interdisciplinair onderzoek. Gelukkig zit dit al geruime tijd in de lift. Denk maar aan de revoluties in velden zoals econofysica en biologische fysica. Dit heeft niet alleen te maken met het feit dat het ene gebied het andere gebied kan aanvullen met zijn expertise en kennis, ook simpelweg een andere invalshoek of manier van redeneren kan zorgen voor nieuwe inzichten, methoden en resultaten.

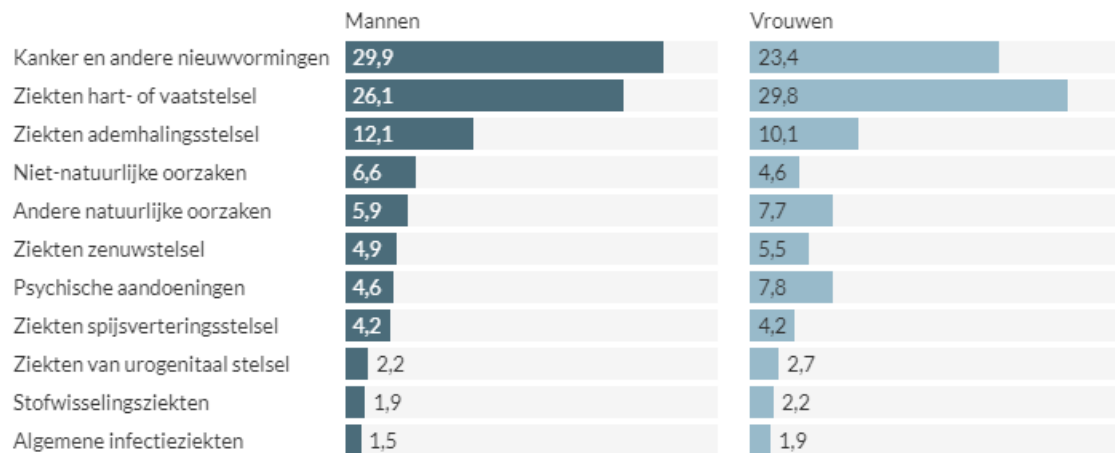
Deze thesis is gesteund op die gedachtegang. Het poogt een fysicus in te zetten om expertise en inzichten uit zijn eigen studiecarière te combineren met de steeds populairder wordende wereld van het machinaal leren om zo het werk van de radiotherapeut, en daarbij de zorg voor de patiënt, verder te optimaliseren.

De rode draad door deze thesis is pas sinds enkele jaren een *hot topic* en is in die tijd het onderwerp geweest van menig onderzoek, iets wat duidelijk zal worden uit de bespreking in sectie 1.4. Deze thesis dient enerzijds om aan te tonen dat het huidige onderzoek veel meer aandacht verdient door te bewijzen dat het een haalbaar doel is en er een massa aan data ter beschikking is, schreeuwend om gebruikt te worden. Anderzijds tracht het ook verdere verfijningen en ideeën voor te stellen en, waar kan, uit te werken.

Conform de modaliteiten is er ook een aspect “wetenschapspopularisering”. Normaal gezien ging dit een “TEDx”-achtige talk zijn georganiseerd in samenwerking met de Vereniging Voor Natuurkunde. Dit is, wegens de coronamaatregelen, echter niet kunnen doorgaan. Ter vervanging publiceert de VVN dagelijks een blogpost over een thesisonderwerp. Via [vvn.ugent.be/blog/automatische-segmentatie-van-organen-in-kankertherapie](http://vvn.ugent.be/blog/automatische-segmentatie-van-organen-in-kankertherapie) zal er een blogpost te vinden zijn, bedoeld om de thesis te communiceren naar het brede publiek. Deze blogpost werd ook apart als bijlage ingediend.

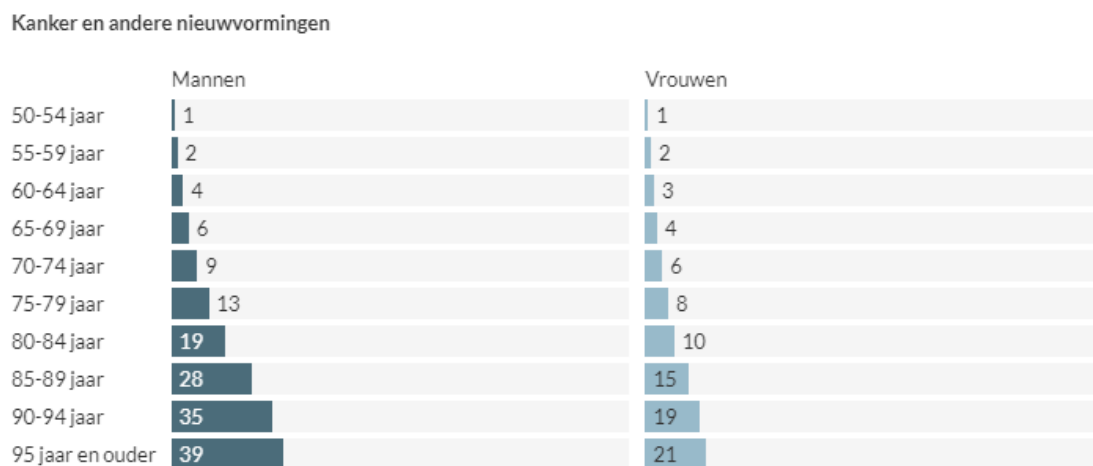
## 1.2 Radiotherapie

Als we kijken naar figuur 1, dan zien we dat kanker de belangrijkste doodsoorzaak is bij mannen en de tweede belangrijkste doodsoorzaak bij vrouwen. Dit heeft verscheidene redenen waarvan we er enkele zullen oplijsten. De meest voor de hand liggende reden is dat men steeds beter wordt in het bestrijden van de andere doodsoorzaken. Denk maar aan de ontdekking van penicilline in 1928, een toen revolutionair antibioticum.



Figuur 1: Relatieve verdeling doodsoorzaken naar geslacht in het Vlaams Gewest in het jaar 2017 [1].

Antibiotica lijken vandaag iets triviaal maar zorgden wel voor een heuse doorbraak in de medische sector. Men wordt ook steeds ouder en kanker is nu eenmaal een ziekte, aanvullend een verzameling van ziekten met meer dan 100 varianten, waarvoor de vatbaarheid toeneemt met de leeftijd (zie figuur 2). Daarnaast kunnen er voor elk van die varianten verschillende oorzaken en behandelingen zijn.



Figuur 2: Aantal sterftegevallen door kanker per leeftijdscategorie per 1000 personen in het Vlaams Gewest in het jaar 2017 [1].

Dit alles maakt van kanker de meest recente en sterkste vijand van de mens in zijn eeuwige strijd tegen de dood. Het gebruik van straling, zowel diagnostisch als curatief, is nog maar courant sinds het begin van de vorige eeuw. In 1895 ontdekte Wilhelm Röntgen de naar hem vernoemde “röntgenstralen”. In 1902 zag William Allen Pusey in deze röntgenstralen een mogelijke behandeling voor Hodgkin lymfoom en andere soorten lymfeklierkanker. Wij zullen focussen op de radiotherapie inzake het curatief verhaal, waarbij ioniserende straling gehanteerd wordt om kankercellen te vernietigen en meer specifiek het DNA van die kankercellen.

Radiotherapie is in staat succesvol tumoren uit te schakelen dankzij twee belangrijke pilaren:

- Het radiobiologische verschil tussen gezonde cellen en kankercellen. Het DNA van gezonde cellen is beter in staat zichzelf te herstellen ten opzichte van het DNA van kankercellen. Het relatief verschil tussen beiden is het grootst bij relatief lage dosissen. Daarom worden bestralingen vaak in kleinere dosissen over meerdere dagen toegediend, in plaats van één grote dosis. Zo kan de tumor stapje voor stapje kapot worden bestraald, maar heeft het gezonde weefsel voldoende tijd tussen twee sessies om te herstellen en geen blijvende schade op te lopen. Doordat radiotherapie gebruik maakt van uitwendige bestraling met hoge energie-fotonen is het niet mogelijk om enkel een dosis toe te dienen aan de tumor en de gezonde weefsels errond volledig te sparen. Om bijwerkingen in de kritische organen (zogenaamde acute en/of late toxiciteit) zo veel mogelijk te beperken, worden er fysische beperkingen opgelegd aan de dosis binnen elk orgaan. Om deze dosis te kunnen evalueren tijdens de radiotherapieplanning aan de hand van dosisberekening per orgaan, moeten de tumor en de kritische organen nauwkeurig gesegmenteerd worden. Dit brengt ons bij het volgende punt.
- Het accuraat intekenen van de kritische organen in de regio van de bestraling en het accuraat intekenen van de tumor. Indien er met een voorafgaand onderzoek een tumor wordt vastgesteld, bestaat de volgende stap erin zo goed mogelijk de tumor en de relevante regio van de patiënt in beeld te brengen aan de hand van bijvoorbeeld een CT-scan. Het is dan enerzijds aan de radiotherapeut-oncoloog om de tumor in te tekenen en anderzijds aan de dosimetrist om de organen in te tekenen. Eens alle relevante gebieden zijn ingetekend, kan er een stralingsplan opgesteld worden, bestaande uit meerdere bundels, met als doel een zo hoog mogelijke dosis te leveren aan de tumor en een zo laag mogelijke dosis aan gezonde weefsels/organen.

Het is nu duidelijk waar autosegmentatie zijn nut heeft. Een CT-scan is een 3D-stack van meestal transversale slices die makkelijk rond de 250 slices (foto's) kan bevatten. Elke foto kan een resem aan organen bevatten zoals het hart, longen, slokdarm,... Deze moeten allemaal zo goed mogelijk worden ingetekend. Dit proces is redelijk tijdrovend en kan persoonsafhankelijk zijn wegens de inter-user variabiliteit van het intekenen. In dat opzicht kan autosegmentatie sterk de tweede pilaar van radiotherapie versnellen. Zo kunnen patiënten niet alleen sneller aan hun behandeling starten, ook hebben de klinici meer tijd over voor andere zaken. Daarnaast brengt het ook een verbetering in de eerste pilaar. Tijdens de behandeling kan de anatomie wijzigen. Bijvoorbeeld doordat de tumor (hopelijk) verkleint of doordat de patiënt gewicht verliest (bijvoorbeeld als gevolg van de chemo). Wanneer het behandelingsplan wordt aangepast op basis van een nieuwe CT scan op een later tijdstip, spreekt men van adaptieve radiotherapie. Om de behandeling te optimaliseren kan men dus na elke behandeling een nieuwe CT-scan maken. Deze behandelingen kunnen zich sneller opvolgen indien het intekenen vlotter kan gebeuren. Daarnaast zullen ze ook doelgerichter (recenter beeld van de tumor) en dus stralingsoptimaler zijn dan wanneer men één CT-scan zou gebruiken voor elke radiotherapeutische behandeling.

## 1.3 Hounsfieldschaal

### 1.3.1 Definitie

Graag spenderen we een subsectie aan de Hounsfieldschaal aangezien onze data (de CT-beelden) met deze schaal werkt. Een goede kennis en goed begrip is daarom vereist om correct de data te kunnen gebruiken. De Hounsfieldschaal is een kwantitatieve schaal gebruikt in CT-beelden die ons iets meer vertelt over de radiodensiteit (zie 1.3.2) van het medium voorgesteld door de pixel. Zo ook is deze waarde gekoppeld aan een bepaald type medium zoals lucht, vet, bot,... Men voert hierbij een lineaire transformatie uit op de originele lineaire attenuatiecoëfficiënt. Dit wordt gedaan zodat de radiodensiteit van gedestilleerd water bij standaard temperatuur en druk gedefinieerd is als 0 Hounsfield-eenheden, terwijl die van lucht gedefinieerd is als -1000. Neemt men nu de gemiddelde lineaire attenuatiecoëfficiënt ( $\mu$ , zie 1.3.3) van een volume-element<sup>1</sup> dan transformeert deze als volgt naar zijn Hounsfield-eenheid (HU):

$$HU = 1000 \cdot \frac{\mu - \mu_{water}}{\mu_{water} - \mu_{lucht}}$$

### 1.3.2 Radiodensiteit

Radiodensiteit is de relatieve onmogelijkheid van X-straling om door een bepaald materiaal te passeren. In dat opzicht is het equivalent aan de opaciteit, maar dan specifiek voor een bepaald deel van het elektromagnetisch spectrum. In het algemeen wordt radiodensiteit gebruikt als een kwalitatieve maateenheid om materialen met elkaar te vergelijken. Het kan kwantitatief gebruikt worden dankzij de Hounsfieldschaal. Iets dat vereist is om correcte (computer)voorspellingen te kunnen maken over hoe sterk en waar de straling zal afnemen tijdens de radiotherapie en daardoor de dosis te kunnen berekenen.

### 1.3.3 Attenuatiecoëfficiënt

Wanneer een bundel fotonen een medium traverseert kan elk foton individueel reageren met het medium. Dit kan via het foto-elektrisch effect (zie 1.3.3.1), comptonverstrooiing (zie 1.3.3.2) en paarvorming (zie 1.3.3.3). Deze interacties hebben allemaal hetzelfde effect: het verminderen van de intensiteit van de oorspronkelijke fotonenbundel. Intensiteit is hier gedefinieerd als het aantal fotonen per oppervlakte-eenheid loodrecht op de bundel.

Neem een dunne laag van een medium. Laat een bundel met initiële intensiteit  $I$  door dit medium gaan. Wanneer deze bundel door de dunne laag met dikte  $\Delta X$  gaat, zal er een vermindering aan fotonen  $-\Delta I$  zijn evenredig met het aantal invallende fotonen  $I$ . Elk foton heeft namelijk dezelfde kans om te interageren.

Zo is er een relatieve vermindering in intensiteit per laagdikte:  $\frac{-\Delta I}{I \Delta X}$ . Dit is een constante per medium<sup>2</sup>. Deze constante vertelt ons de kans dat een foton met een bepaalde energie verdwijnt per eenheid van dikte en wordt de attenuatiecoëfficiënt genoemd. Uit  $\mu = \frac{-\Delta I}{I \Delta X}$  kan men een formule afleiden voor de intensiteit van de mono-energetische energiebundel in functie van de afgelegde afstand.

---

<sup>1</sup>In het geval van CT-scans is dit volume-element een zogenaamde “voxel”. Met afmetingen van de oppervlakte van 1 pixel, en als 3de lengte de snededikte van één CT-beeld.

<sup>2</sup>Afhankelijk van de energie van de fotonenbundel. Elke energie heeft een verschillende kans om te interageren, zie secties 1.3.3.1, 1.3.3.2 en 1.3.3.3.

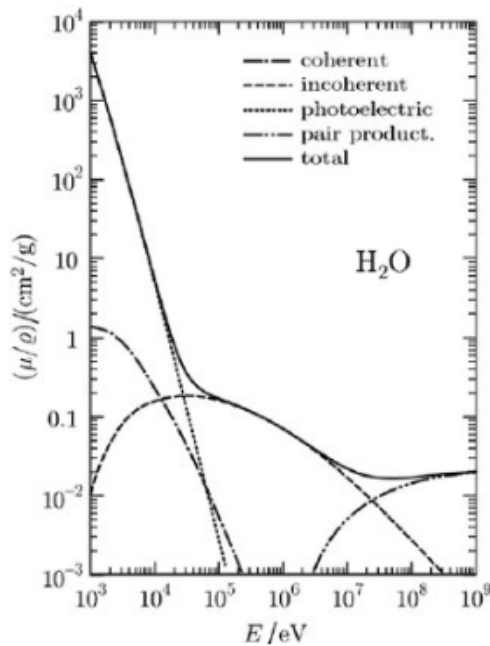
$$\mu = \frac{-\Delta I}{I\Delta X} \Leftrightarrow \Delta I = -\mu\Delta X I$$

Hierin herkent men de typische differentiaalvergelijking van een exponentiële met als oplossing:

$$I(x) = I_0 e^{-\mu x}$$

Met  $x$  de afgelegde afstand in het medium en  $I_0$  de initiële bundelintensiteit.

De attenuatiecoëfficiënt van een stof is recht evenredig met de massadichtheid. Zo krijgt men dat de attenuatiecoëfficiënt van longen slechts een derde is vergeleken met die van spierweefsel. Met deze verschillen in attenuatie van de fotonenbundel kan men dus aan radiologische beeldvorming doen.



Figuur 3: Verloop van de attenuatiecoëfficiënt in functie van de energie voor de drie verschillende types interacties en de totale som aan attenuatie. Coherent en incoherent vallen alletwee onder “Comptonverstrooiing” [2].

X-stralen kunnen via drie verschillende mechanismen interageren met materie. Elke interactie heeft zijn eigen type materie- en energieverloop. Het totale attenuatieverloop in functie van energie is dan een som van deze drie mechanismen (zie figuur 3). Deze drie mechanismen zullen nu uitgelegd worden.

### 1.3.3.1 Foto-elektrisch effect

Bij het foto-elektrisch effect wordt het foton geabsorbeerd door een zwak gebonden elektron. Hiermee verdwijnt het foton en ontsnapt het elektron aan zijn atoom. Aangezien vooral elektronen uit de K- en L-schillen het slachtoffer zijn, blijft er achteraf een vacature over. Deze vacatures zullen ingevuld worden door hoger gelegen elektronen, met uitzending van karakteristieke X-straling als gevolg. Deze secundaire straling speelt geen rol bij medische beeldvorming aangezien deze van lage energie is en bijgevolg volledig geabsorbeerd wordt door de patiënt. De kans op dit fenomeen schaalst met  $Z^5$  (atoomgetal  $Z$ ) en met  $E_{foton}^{2/7}$  [3]. Dit fenomeen speelt dus vooral een rol bij lage energieën en materialen met een groot atoomgetal.

### 1.3.3.2 Comptonverstrooiing

Bij deze interactie “botst” een foton met een orbitaal elektron van een atoom. Hierbij geeft het foton slechts een deel van zijn energie af aan het elektron en beide deeltjes reizen verder. De energie van het verstrooide elektron is afhankelijk van de invalshoek en van de energie van het invallend foton en is als volgt:

$$E_{foton}^{vers} = \frac{E_{foton}^{inval}}{1 + \frac{E_{foton}^{inval}}{0.511}(1 - \cos(\theta))}$$

Met de energie in MeV (0.511 MeV is de rustenergie van het elektron). De kans op dit fenomeen schaalst met  $Z$  (atoomgetal  $Z$ ) en met  $E_{foton}^{1/2}$ . Door deze geringere afname in functie van energie zal dit proces domineren over het foto-elektrisch effect bij hogere energieën en dan vooral bij X-stralendiagnostiek. Dit type straling krijgt dan ook veel aandacht aangezien de secundaire straling niet per se intern geabsorbeerd wordt en de verstrooide fotonen voor afname van beeldkwaliteit zorgen. Daarnaast kan ze ook stralingsbelastend zijn voor de practicus.

### 1.3.3.3 Paarvorming

Paarvorming kan zich voordoen bij zeer hoog energetische fotonen. Een dergelijk foton kan, in de buurt van een atoomkern, omgezet worden in een elektron-positron paar. De drempelwaarde is 1.022 MeV (twee keer de rustenergie van een elektron en zo ook van een positron). Deze energieën spelen geen rol bij X-stralen voor medische beeldvorming en zullen algemeen pas echt dominant worden bij energieën vanaf 20 MeV.

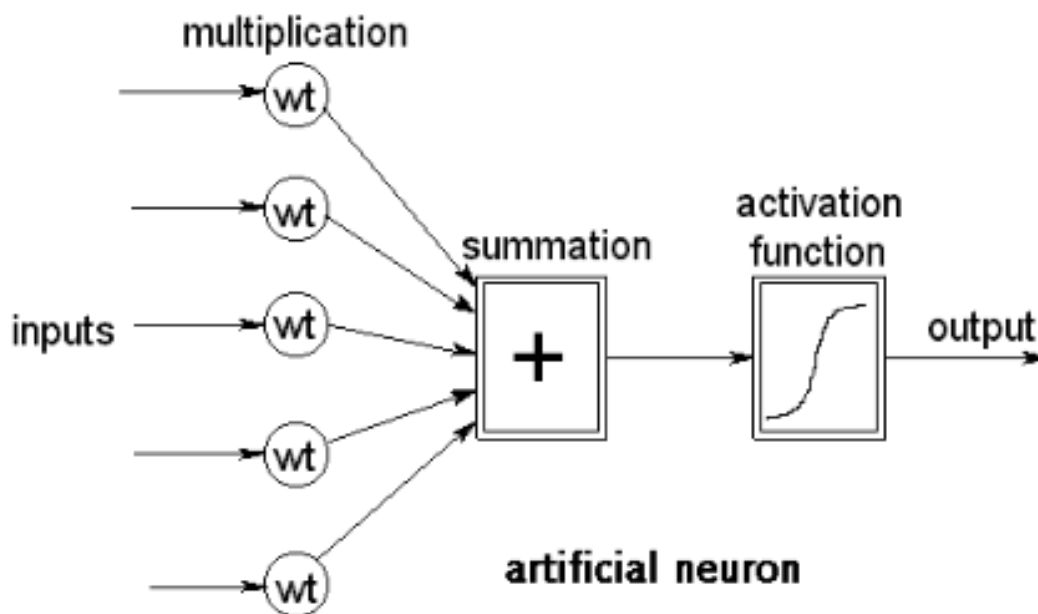


## 1.4 Deep learning in de medische wereld

### 1.4.1 Een neurale netwerk

Machine learning is het trainen van een programma om beter te worden in een bepaalde taak door ervaring op te doen in het uitvoeren van deze taak. In dat opzicht kan men het vergelijken met een mens die leert fietsen. Hoe meer men oefent en (in)directe feedback krijgt, hoe beter de persoon zal worden. Uiteindelijk bereikt het leren een einddoel en is de mens/het computerprogramma in staat om te fietsen/een bepaalde taak uit te voeren.

In figuur 4 wordt een neuron visueel voorgesteld. In deze voorstelling krijgt het neuron vijf inputwaarden, die elk worden vermenigvuldigd met een trainbaar gewicht. Na het sommeren wordt er nog een trainbare bias aan toegevoegd en als laatste gaat dit getal door een activatiefunctie. Indien een model gebruik maakt van slechts één neuron wordt dit een perceptron genoemd. Een neuron kan op zichzelf alleen aan binaire classificatie doen. Men kan meerdere neuronen samenbrengen in een laag om complexere taken te kunnen uitvoeren. In theorie zou één laag met een zeker aantal neuronen moeten volstaan om eender welke taak uit te voeren dat een complexer netwerk ook kan. Echter kunnen diepere netwerken het probleem vaak efficiënter aanpakken met minder te trainen parameters in totaal. Ze kunnen complexere functies voorstellen met minder “hardware”. Een diep neurale netwerk is een netwerk met één of meerdere lagen tussen de input- en outputlaag. Indien elk neuron van een laag verbonden is met elk neuron van de vorige en volgende laag spreken we van een volledig verbonden neurale netwerk.

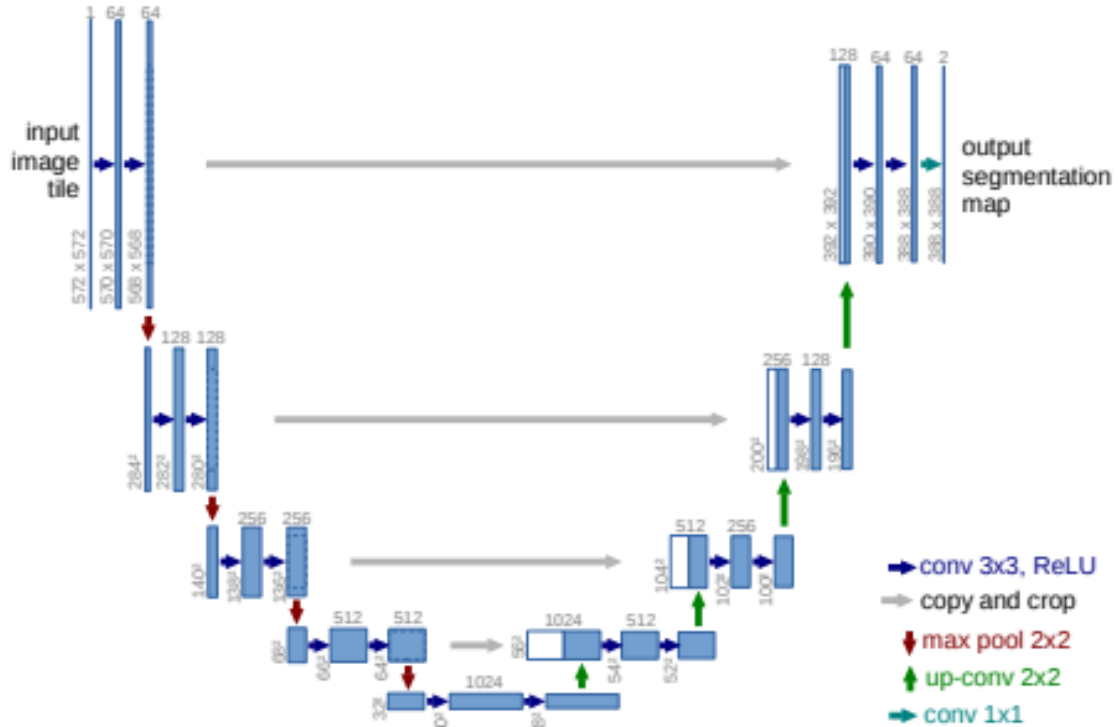


Figuur 4: Visualisatie van de werking van een artificieel neuron.

Het ontwerpen van een neurale netwerk dat niet te complex is maar wel nog de gewenste taak kan uitvoeren, is een studie en kunst op zich. Voor dit onderzoek hebben we zelf geen netwerk gemaakt, maar gebruiken we een architectuur die zichzelf al heeft bewezen in het segmenteren van afbeeldingen: U-Net.

## 1.4.2 U-Net

U-Net is een diep convolutioneel neuronaal netwerk dat al meermaals zichzelf heeft bewezen in het kunnen segmenteren van afbeeldingen [4] en waarvan de architectuur is weergegeven in figuur 5.



Figuur 5: U-net architectuur voor een afbeelding van 572x572 pixels (Ronneberger, Fischer and Brox, 2020).

Links komt er 1 afbeelding binnen, rechts komen er zoveel afbeeldingen uit als verschillende categorieën die men wil labelen in de oorspronkelijke afbeelding. Het netwerk zelf bestaat uit een contracterend pad links en een expansief pad rechts. Het contracterend pad heeft de typische vorm van een convolutioneel netwerk waarbij herhaaldelijk twee 3x3 convoluties worden opgevolgd door ReLU activatiefuncties en 2x2 max pooling (downsampling) operaties. In dat opzicht leert de linkerkant vooral “high level features” zoals “hoe herken ik een verticale lijn” of “hoe vormen deze krommen samen een ovaal”. Deze kant zou gebruikt kunnen worden om CT-beelden te classificeren als “dit heeft een long” of “dit heeft een hart”. U-Net gaat nog een stapje verder en gebruikt het expansief pad om effectief de long en het hart aan te duiden in de afbeelding. In dat opzicht leert het meer “low level features”. Dit doet het door enerzijds, per downsampling, ook een upsampling uit te voeren. Waarbij er ook geconcateneerd wordt met de desbetreffende featuremaps uit het linkse pad. Op het einde worden nog 1x1 convolutiefilters gebruikt om de 64 lagen om te vormen naar het gewenste aantal lagen gelijk aan het aantal categorieën.

### 1.4.3 Gerelateerde onderzoeken

Hoewel een orgaan een 3D-structuur is en dus meerdere CT slices beslaat, heeft automatische segmentatie op basis van 2D beelden reeds resultaten opgeleverd [5]. Zhou *et al.* gebruikten slechts 240 CT-beelden om 19 anatomische structuren in de thorax te leren herkennen. Waarbij als finaal resultaat een test IoU (zie 2.4.2 voor definitie IoU) score van 87.9% IoU werd bekomen. Dit is niet zo absurd aangezien de clinici die de organen annoteren feitelijk ook in 2D werken. Zij bekijken slice per slice welke organen waar zitten. Natuurlijk kan de radiotherapeut-oncoloog ook contextuele informatie gebruiken zoals “het hart moet tussen de longen liggen”. Dit is echter lokale 2D-informatie en dit kan het netwerk ook leren.

Het trainen van een netwerk kan geruime tijd in beslag nemen, soms zelfs enkele dagen. In sectie 1.4.2 zagen we dat de linkerkant van het netwerk de typische vorm heeft van een convolutioneel netwerk en dat deze kant vooral bedoeld is om “high level features” te leren. Men kan de redenering maken dat een cirkel een cirkel is, of deze nu deel is van een luchtpijp of van een koplamp. Daarom bestaat er een online dataset waarbij men convolutionele netwerken en hun gewichten kan downloaden genaamd “ImageNet”<sup>3</sup>. Dit deel van het U-Net netwerk wordt ook wel de “backbone” genoemd. Ter verduidelijking: feitelijk zou elk netwerk op zichzelf trainbaar moeten zijn. Deze aanpak kan het trainproces echter wel versnellen. Het eerder aangehaalde onderzoek had na 80,000 epochs<sup>4</sup> nog steeds geen getraind netwerk. Dit lukte pas na 160,000 epochs. Terwijl het gebruiken van de voorgetrainde gewichten al een getraind netwerk leverde na 22,000 epochs.

Een belangrijke keuze is dan welke backbone er gebruikt zal worden. De VGG16 structuur heeft zichzelf al bewezen als een goeie afweging tussen performantie en trainingstijd [6], alsook als een backbone waar ImageNet veel ervaring mee heeft. Deze backbone genoot dan ook de voorkeur en werd gebruikt in het trainen van ons eigen netwerk.

Eerder gepubliceerde resultaten met betrekking tot autosegmentatie zullen gebruikt worden om onze eigen resultaten aan te toetsen. We vermelden hier de resultaten van drie onderzoeken waarvan onderzoek 1 de IoU score (sectie 2.4.2) gebruikte [5] en onderzoek 2 en 3 de Dicescore (sectie 2.4.1) gebruikten [7, 8]. Voor het laatste onderzoek gebruiken we de resultaten van hun U-net gebaseerd model.

Tabel 1: Behaalde resultaten van gelijkaardige onderzoeken.

Orgaan	Onderzoek 1 (IoU)	Onderzoek 2 (Dice)	Onderzoek 3 (Dice)
Hart	0.817	0.941	0.85
Slokdarm	0.107	0.858	0.71
Longen	0.903	N/A	0.965
Luchtpijp	N/A	0.926	N/A
Ruggenmerg	N/A	N/A	0.83

<sup>3</sup><http://www.image-net.org/>

<sup>4</sup>Een epoch is gedefinieerd als eenmaal volledig over de trainingsset gaan.

Als laatste willen we aanhalen dat het creëren van een perfect model vermoedelijk niet mogelijk is. Dit is ten gevolge van het niet perfect zijn van de mens. Hoe een radiotherapeut-oncoloog een orgaan intekent kan afhangen van persoon tot persoon. Aangezien ons model zal leren van geannoteerde afbeeldingen van verschillende radiotherapeut-oncologen, zal het geen uniforme informatie krijgen. Daarbij komend zullen we ook niet kunnen bepalen of ons model “perfect” is, omdat we het meten aan niet 100 % perfecte data. Tabel 2 vertelt ons welke verschillen er kunnen zijn tussen enkele thoracale organen. In totaal bestudeerde men 198 structuren gecontournd door 21 klinici. De uniformiteit tussen verschillende klinici wordt gegeven in termen van hun Dice (sectie 2.4.1) en IoU (sectie 2.4.2) score [9].

Tabel 2: Verschillen in ingetekende contouren tussen verschillende klinici.

Orgaan	IoU score	Dicescore
Hart	0.86	0.92
Slokdarm	0.48	0.64
Longen	0.95	0.97
Ruggenmerg	0.60	0.70

Hieruit leren we dat er alsnog sterke verschillen kunnen zijn tussen het intekenen van de contouren. Wegens deze intrinsieke verschillen gaan wij het hoogstwaarschijnlijk niet beter kunnen doen dan deze scores. We zullen dus niet streven naar een perfecte score, maar naar een score evenwaardig met de verschillen tussen klinici. Daarbijkomend zullen we extra visuele en andere inspecties moeten uitvoeren om zeker te kunnen zijn dat onze fouten binnen de klinische variaties vallen en niet ten gevolge zijn van een slecht model.

## 1.5 Doel en probleemstelling

Als laatste lichten we concreet toe wat wij willen bereiken. Waarin gelijken we op vorige onderzoeken en waar wijken wij af? Wat willen we exact dat het model kan en waar verwachten we moeilijkheden?

### 1.5.1 Doel

In essentie willen wij hetzelfde als de onderzoeken aangehaald in sectie 1.4.3 en omvat in de titel van deze thesis: orgaansegmentatie met behulp van deep learning. Deze segmentatie zal getoetst worden aan de resultaten van de onderzoeken in 1.4.3 om te zien of ze competitief zijn. We mogen ook niet vergeten dat 100% accuraatheid geen logisch, laat staan realistisch, streefdoel is. Ook tussen het intekenen van organen door verschillende radiotherapeuten-oncologen zal er een verschil zitten. Naast deze doelstelling gaan wij nog een stap verder. We willen ook een pipeline creëren om deze autogesegmenteerde beelden terug in een DICOM RT structfile<sup>5</sup> te steken. Deze kunnen dan gebruikt worden om de autogesegmenteerde modellen in Raystation<sup>6</sup> in te laden.

### 1.5.2 Organen en tumor

Er moet ergens een keuze gemaakt worden welke en hoeveel organen we willen trainen. De keuze maken voor te weinig en/of te gemakkelijk en we riskeren aan “overkill” te doen waarbij we een diep neurale netwerk gebruiken voor een taak waar veel eenvoudigere methoden voor gebruikt kunnen worden. Kiezen voor te veel en/of te complex en we riskeren amper resultaten te kunnen krijgen waardoor we niet kunnen weten of het aan de grote hoeveelheid (complexe) organen ligt of aan het model. Met dit in het achterhoofd worden er vijf organen en de tumor zelf meegenomen. Een motivatie en “moeilijkheidsgraad” van elk van de organen en de tumor wordt hieronder gegeven.

#### 1.5.2.1 Hart

Het hart is het orgaan dat vermoedelijk het minst voorkomt op alle CT-beelden (lees: het kleinste aantal slices inneemt) maar wel vaak een redelijk groot oppervlakte inneemt. Het gaat redelijk goed op in zijn achtergrond maar is nog te herkennen wegens een dunne omranding die er net wat anders uitziet. Dit orgaan heeft dus het nadeel dat het niet sterk in het oog springt maar het voordeel dat het zeer groot is. Vermoedelijk zal het model wel in staat zijn dit orgaan te leren herkennen.

#### 1.5.2.2 Slokdarm

De slokdarm beslaat een groot aantal van de CT-beelden. Het is een zogenaamd “serieel” orgaan. Dit betekent dat het ernstig beschadigen van het orgaan op één plaats nadelige gevolgen kan hebben voor het functioneren van het gehele orgaan. De slokdarm is relatief dun en gaat gemakkelijk op in zijn achtergrond. Vermoedelijk zal het model meer moeite hebben met dit orgaan te leren herkennen.

---

<sup>5</sup>DICOM (RT) is een wereldwijde standaard voor uitwisseling van medische gegevens (die specifiek betrekking hebben op RT). Zo goed als alle medische commerciële softwarepakketten (waaronder Raystation) bevatten DICOM. Deze zijn nodig om de segmentaties op te slaan.

<sup>6</sup>Raystation is een radiotherapieplanningsoftware die gebruikt wordt voor het segmenteren van de CT-beelden tot het plannen en optimaliseren van de stralingsbehandeling

### 1.5.2.3 Tumor/GTV

Het leren herkennen van de tumor (en meer bepaald het GTV<sup>7</sup>) is fundamenteel anders dan het leren herkennen van de organen. Er is geen verwachte locatie voor dit weefsel. Er valt niet makkelijk predicatief te leren. In dat opzicht is het leren herkennen meer een outlier detectie probleem. Het model zou moeten leren wat er zou moeten zijn, en dan een alarm geven als het dat niet ziet. We verwachten niet veel van dit weefsel maar nemen dit mee indien het model uitzonderlijk goed blijkt te zijn in het herkennen van de organen om te weten of het model misschien meer aankan dan we initieel verwachtten. De variatie binnen de tumoren is ook zeer groot. De positie en de grootte kunnen sterk verschillen. Sommige tumoren, zoals tumoren die centraal in een long zitten, zijn beter afgelijnd dan tumoren die dicht bij het mediastinum zitten en daar deels mee vergroeid zijn.

### 1.5.2.4 Longen

De longen zijn veruit de makkelijkste organen om te voorspellen. Ze hebben het voordeel dat ze zowel uit een zeer verschillend/verschillende medium/dichtheid bestaan en ook nog eens ontzettend veel oppervlakte innemen op de CT-beelden. Een simpele threshold op de Hounsfieldunits zou al een manier zijn om het merendeel van de longen (maar ook de luchtpijp) te identificeren. Dit orgaan is zo eenvoudig dat bij het annoteren door de radiotherapeut-oncoloog al een machinale techniek gebruikt wordt. Men duidt een punt aan binnen een long en dit gebied breidt zich uit tot het een duidelijke omslag in de dichtheid vindt. In dat opzicht zullen de longen meer dienen als een indicator om te testen of het model wel zaken kan trainen en voorspellen. Indien het faalt in het voorspellen van de longen zal het zeker niet in staat zijn de andere organen te leren herkennen.

### 1.5.2.5 Luchtpijp

De luchtpijp is opnieuw een serieel orgaan. Het heeft echter het sterke voordeel van een zeer verschillend type medium/weefsel te zijn ten opzichte van de andere organen (buiten de longen). Het grote verschil met de longen is de oppervlakte in de doorsneden en de positie. We verwachten geen moeilijkheden met dit orgaan, maar denken niet dat het even goed zal zijn als het autosegmenteren van de longen.

### 1.5.2.6 Ruggenmerg

Net zoals de luchtpijp en de slokdarm is dit een orgaan dat door veel CT-beelden gaat. Ook dit orgaan is een zogenaamd serieel orgaan. Het is de meest geannoteerde categorie van de zes besproken categorieën. Het orgaan zelf is echter redelijk klein en op zichzelf niet makkelijk terug te vinden. Het heeft wel het voordeel consistent een zekere omringing van bot te hebben wat het alloceren door het model zou moeten bevorderen. We verwachten dat dit model ongeveer even moeilijk of makkelijk als de slokdarm zal zijn.

---

<sup>7</sup>GTV staat voor Gross Tumour Volume. Dit omvat de tumor en eventueel wat oedeem dat niet meteen duidelijk te herkennen is op de CT-beelden.

## 2 Methoden

Elk onderzoek of experiment heeft in sé hetzelfde stappenplan. Het uitdenken van de proef, het voorbereiden van het specimen, het specimen onderwerpen aan het experiment en het valideren van de resultaten. Deze sectie zal het stappenplan voor onze experimenten uitlijnen. Hoe bereiden we onze specimens voor, of concreet: wat moet er met onze data gebeuren vooraleer het model er op kan trainen? Op welke manieren proberen we ons model te trainen om de taak correct uit te voeren? Hoe kwantificeren we het al dan niet correct uitvoeren van de taak en welke experimenten zullen we uitvoeren?

### 2.1 Datavoorbereiding

Datavoorbereiding is 90 % van het experiment. Zonder deftige data hebben we niks om op te trainen. Daarnaast kan het correct voorbereiden van de data het experiment maken of kraken. De “garbage in, garbage out” doctrine leert ons dat we alleen zinnige resultaten kunnen bekomen indien we zinnige input leveren. Daarom moet er voldoende aandacht gespendeerd worden aan welke stappen we ondernemen om de ruwe data klaar te stomen voor het trainen van het model.

#### 2.1.1 Masker

Bij het intekenen duiden de radiotherapeut-oncoloog (voor de tumor) en een dosimetrist (voor de “organs at risk”) een reeks van punten aan die samen een gesloten contour vormen. Deze vorm van contourweergave moet vertaald worden naar een “masker” dat een netwerk snel en eenvoudig kan gebruiken. Er werd gekozen om een pixel te “activeren”<sup>8</sup> indien het centrum van de pixel binnen deze contour ligt. Dit is natuurlijk geen exacte vertaling van het getekende orgaan (zie afbeelding 6), maar is binnen de praktische doeleinden goed genoeg.

Elk CT-beeld is opgebouwd uit 512 x 512 pixels. Elke pixel heeft een lengte en hoogte van 1.383 mm. Dit betekent dat de oppervlakte 1.913 mm<sup>2</sup> bedraagt. Gezien de algemeen concave vorm van de (getekende) organen mogen we aannemen dat de grootste fout zich zal voordoen bij een rechte door de pixel die net niet of net wel zorgt voor een activatie. De fout op de oppervlakte van de geactiveerde pixel kan hierbij niet groter zijn dan  $\frac{1.913\text{mm}^2}{2} - \epsilon$ . Deze fout is kleiner dan de fout bij het afstellen van de besturingsapparatuur [10].

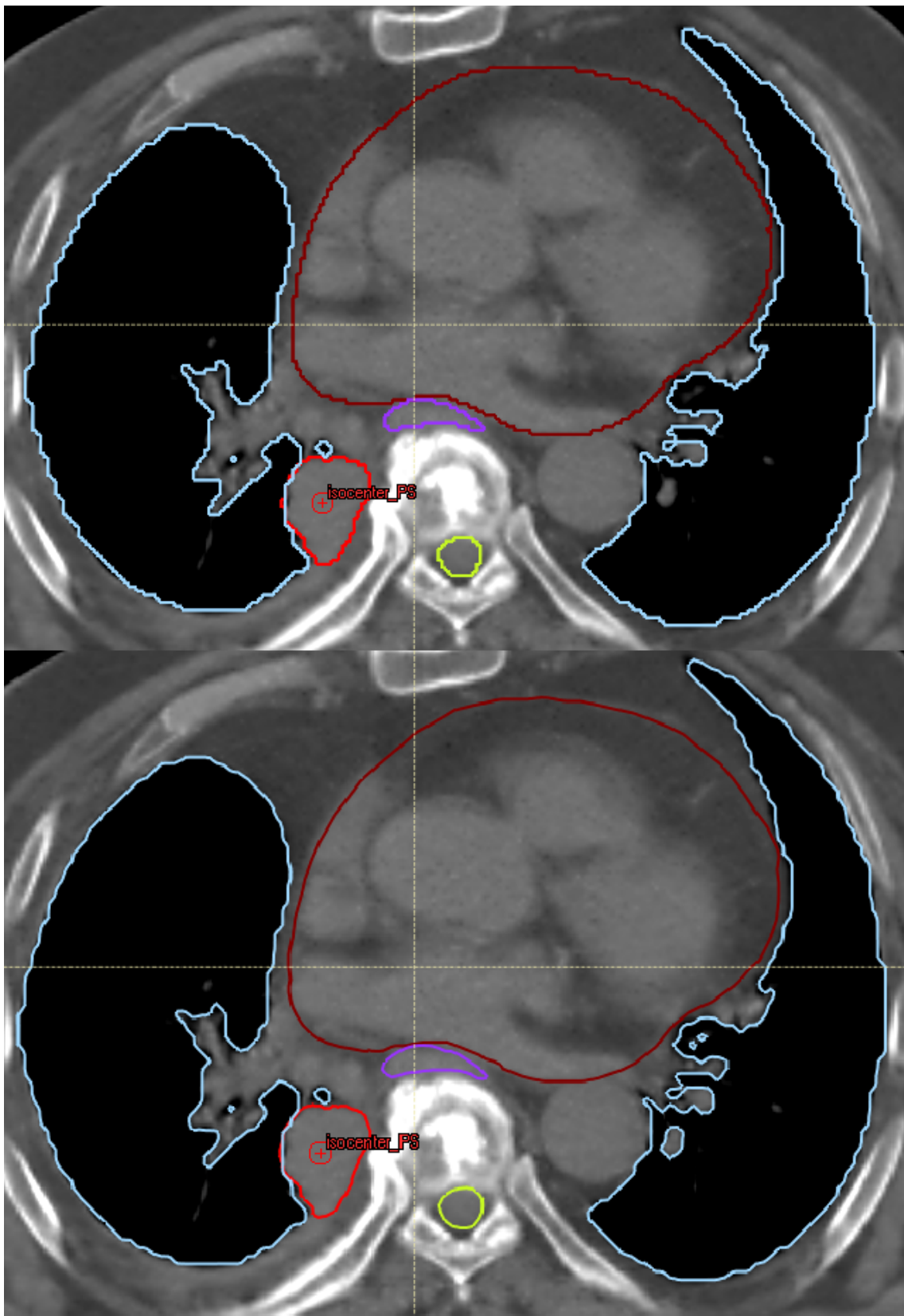
In deze studie werken we met tumoren in de thorax regio. Een patiënt tijdens de behandeling die net wat anders heeft gegeten dan bij de diagnose, of simpelweg het ademen van de patiënt, zal een fout introduceren groter dan  $\frac{1.383\text{mm}}{2} - \epsilon$ .

Daarnaast hebben we nog de verschillen tussen klinici onderling die groter zijn dan onze geïnduceerde verschillen (zie tabel 2).

Dit alles bij elkaar bewijst dat het activeren van een pixel indien het centrum in de contour ligt voldoende nauwkeurig is.

---

<sup>8</sup>Het labelen van deze pixel als behorend tot het orgaan.



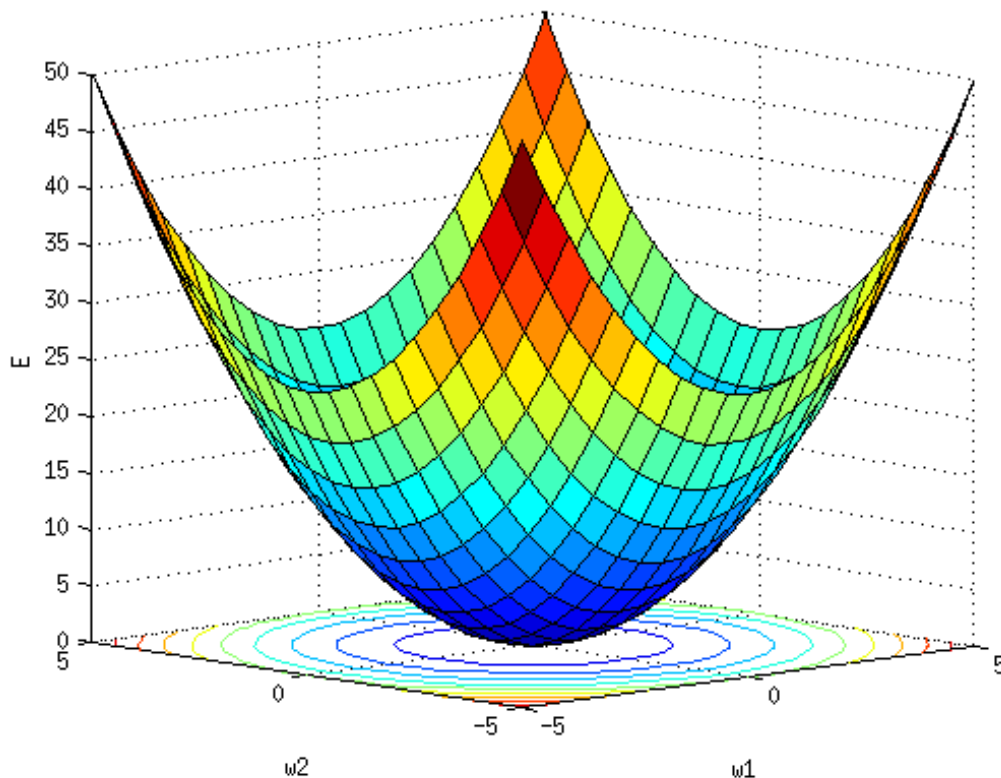
Figuur 6: De getekende contouren (onder) en hun pixelwise equivalenten (boven). Donkerrood: hart, rood: GTV (tumor), blauw: longen, paars: slokdarm, geelgroen: ruggenmerg.



### 2.1.2 Datanormalisatie

Een invoer (van bijvoorbeeld een pixelwaarde) in het netwerk zal vermenigvuldigd en opgeteld worden met verschillende trainbare parameters. Hoe deze parameters worden getraind en aangepast, wordt bepaald door de output van het netwerk en de werkelijke gewenste waarde. Het aanpassen van deze parameters is echter niet afhankelijk van slechts één input, maar van meerdere inputs. Hoe drastisch de trainbare parameters veranderen, wordt dus vooral bepaald door de berekende output die het sterkst verschilt van de werkelijke waarde. Daarom is het belangrijk dat het bereik van alle pixelwaarden genormaliseerd is zodat elke pixel ongeveer even veel bijdraagt.

Een ander voordeel van normalisatie is dat “gradient descent” sneller convergeert als men normalisatie toepast dan wanneer men dit niet doet. Gradient descent is een techniek waarbij men stappen neemt proportioneel aan de negatieve van de afgeleide van de functie die men evalueert. Deze functie is bij het trainen van een netwerk de zogenaamde “loss” functie. Het schalen van de inputwaarden zorgt ervoor dat de oppervlakte van de zogenaamde “error surface” (zie figuur 7) een meer sferische vorm krijgt, terwijl het anders een nogal sterk krommende ellipsoïde zou zijn. Aangezien gradient descent krommingonwetend is, zou een error surface met een sterke kromming ervoor zorgen dat men veel onnodige stappen neemt. Wanneer we de inputwaarden schalen, verminderen we de kromming, wat methoden die kromming negeren (zoals gradient descent) beter doen werken. Sterker nog: wanneer de errorsurface sferisch is, wijst de afgeleide recht naar het minimum. Dit zorgt ervoor dat het programma beter leert.



Figuur 7: Voorbeeld van een “error surface”. Hier met slechts twee trainbare parameters “w1” en “w2” voor eenvoudigere visualisatie. Op de verticale as een maateenheid voor de fout of “loss” [11].

Er zijn verschillende herschalingstechnieken. De techniek die we kiezen wordt mee bepaald door de vorm van de data. In dit onderzoek is het belangrijk te beseffen dat de data al een zekere standaardisering heeft ondergaan. De pixelwaarde is al omgerekend naar zijn zogenaamde “Hounsfieldwaarde” (meer uitleg in sectie 1.3). Wat we nu al horen te weten, is dat het bereik van Hounsfieldwaarden in CT-beelden gaat van -1000 tot en met 3095 [12].

Het is belangrijk dat de info die de Hounsfieldwaarden ons geeft, bewaard blijft in onze herschaling van de data. Met andere woorden dat elke pixel van elke mogelijke input dezelfde transformatie ondergaat. Met dit in het achterhoofd en het feit dat er al een intensiteitsbereik bekend is, is het optimaal om min-max normalisatie toe te passen waarbij het interval [-1000, 3095] wordt geprojecteerd op het interval [0, 1]. Op die manier wordt het zwaartepunt van de data in een genormaliseerd bereik geplaatst. Dit gebeurt door onderstaande formule:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Waarbij  $x$  de oorspronkelijke pixelwaarde is,  $x'$  de herschaalde,  $\min(x) = -1000$  en  $\max(x) = 3095$ . We kunnen makkelijk nagaan dat dit -1000 transformeert naar 0, en 3095 transformeert naar 1.

De aandachtige lezer ziet meteen de gelijkenis met de min-max normalisatie uit sectie 1.3. Dit komt omdat er in essentie hetzelfde principe gebeurt. We voeren een lineaire transformatie uit op een bereik aan waarden, waarbij we de transformatie van 2 punten vooraf vastleggen die dan de volledige bijectie bepalen.

Een belangrijke opmerking die we hier moeten maken, is dat niet elke pixel in het bereik [-1000, 3095] viel. Van de eerste 29 patiënten (goed voor iets meer dan 2 miljard pixels) was er 0,0064 procent die lager was dan -1000 en 0,00369 procent die hoger was dan 3095. Samen is dit goed voor 0,01 procent van alle pixels die niet naar [0, 1] getransformeerd werden. Dit is echter geen probleem aangezien, zoals aangehaald, het zwaartepunt een goede normalisatie heeft gehad en deze normalisatie gestaafd is op het verwachte bereik. De 0,01 procent aan “outliers” zal dus minder kwaad doen dan wat de normalisatie aan goeds doet. Die 0,01 procent is ook grotendeels te verklaren door bronnen zoals in figuur 8. De vier donkerpaarse regio’s buiten de ronde CT-foto zijn geen deel van het CT-beeld maar is pure “filler”. Deze pixels hebben standaard pixelwaarde -2048. Deze ondergrens werd niet meegenomen in het normalisatiebereik aangezien ze geen deel is van de fysisch/biologisch interessante regio. Daarnaast is er ook een geelgroene regio die helder oplicht. Dergelijke hoge Hounsfield waarden zijn afkomstig van vaste stoffen zoals steen, al lijkt een metalen implantaat logischer. In het geval van de foto was het een schouderprothese. Opnieuw zijn dit soort pixels niet deel van de fysisch/biologisch interessante regio aangezien ze niet gebruikt kunnen en/of mogen worden voor het trainen van het model.



Figuur 8: CT-beeld uit de eigen dataset met een bron van pixels  $< -1000$  en een bron van pixels  $> 3095$ .

### 2.1.3 Data-augmentatie

Een finale en belangrijke stap is het augmenteren van de data tijdens het trainen. Concreet worden er realistische kleine veranderingen toegepast op de traindata. Dit verbetert de robuustheid van het model aangezien het overtrainen tegengaat. Ook creëren we zo artificieel veel meer data. Data-augmentatie is best niet te sterk, anders zijn de beelden te onrealistisch en kan er ook niet geconvergeerd worden. Daarnaast is een te zwakke data-augmentatie ook niet wenselijk want dan kunnen we evengoed niet augmenteren. We kozen voor het kunnen roteren van de afbeeldingen over  $[-5, 5]$  graden en het kunnen schalen over een factor  $[0.9, 1.1]$ . Dit wegens het succes dat het had in een eerder gelijkaardig probleem [13].

Aangezien de organen zich vooral in het centrum van de foto bevinden en er dus een hoop dode informatie in de foto is, werd er voor sommige testen gekozen om enkel de centrale  $256 \times 256$  pixels te gebruiken. In dat geval moeten de gewichten vermeld in secties 2.2.2 & 2.2.3 uiteraard herschaald worden volgens het principe dat alle pixels buiten de centrale  $256 \times 256$  regio “niet orgaan” zijn.

### 2.1.4 De HPC

De HPC is de “High Performance Computing Infrastructure” van de UGent. Onze laptop heeft slechts een Quad-Core met 1.60 GHz, 16 GB RAM en 2 GB GPU, terwijl de HPC onder andere een cluster met 10 nodes heeft. Elke node heeft 256 GB RAM,  $2 \times 16$  cores met 2.8 GHz en  $4 \times 32$  GB GPU. Gezien een GPU echt wel een vereiste is om vlot met foto’s te kunnen trainen en al bij al onze hardware niet aan de HPC kan tippen werd deze infrastructuur dan ook gebruikt.

Een nadeel is natuurlijk dat deze infrastructuur gedeeld is en dat het soms een tijd kan duren tegen dat een “job” (lees: een code die men indient om te runnen) wordt aanvaard en gerund. Daarom werd voor kleinere testen Google Notebooks gebruikt. Hierbij kan men onder andere genieten van 12uur 13GB RAM per 24u per account. Wat betekent dat als men een tweede Gmail-account aanmaakt en de workbook daarmee deelt, men een effectieve GPU-tijd kan krijgen van 24u. Lange jobs (’s nachts) runnen is helaas niet mogelijk aangezien er een cut-off is na 90 minuten inactieve tijd.

### 2.1.5 Geshuffelde schijnpatiënten

Kijkende naar de hardwarelimieten van de HPC werd een batchgrootte van acht gebruikt. Zo hadden we een minimum aan GPU’s en nodes nodig waardoor er vaak getraind kon worden. Dit heeft twee nadelen indien we patiënt per patiënt foto’s zouden inladen. Enerzijds zullen de acht foto’s grotendeels op elkaar lijken aangezien de foto’s per patiënt grotendeels geordend zijn<sup>9</sup>. Daarnaast zullen enkele batches na elkaar foto’s van dezelfde patiënt meegegeven worden. Rekening houdend met het feit dat de gewichten van het model worden aangepast na elke batch zien we dat op deze manier de gewichten niet helemaal efficiënt getraind worden. Ze worden enerzijds te hard in een bepaalde richting geduwd na elke batch omdat die batch ongeveer dezelfde foto’s bevat. Anderzijds worden ze ook te hard in een globale richting geduwd per patiënt aangezien er ook verschillen zijn tussen verschillende patiënten.

Dit alles zorgt ervoor dat de loss trager zal dalen dan gewenst is en daar bijkomend kleine pieken kan hebben na het doorlopen van een patiënt.

Om die reden werden er zogenaamde “schijnpatiënten” gecreëerd. Neem nu opnieuw onze eerste 29 patiënten (waarbij elke patiënt rond de 250 foto’s telt). Deze werden opgedeeld in drie groepen met grootte 19, 5 en 5 voor respectievelijk train, validatie en test<sup>10</sup>. Dit is om zeker geen contaminatie te hebben tussen deze drie groepen. Binnen elke groep werden alle foto’s door elkaar gehaald en terug opgeslagen in een hoeveelheid “schijnpatiënten”. Opslaan in één groot bestand per groep was niet wenselijk wegens enerzijds de grootte en anderzijds de behoefte om soms te testen op een kleiner aantal patiënten.

We vermelden graag ook dat na elke trainepoch de lijst met de schijnpatiënten werd geshuffeld. Anders kunnen we ook een piek in de losswaarden verwachten na elke epoch omdat het model zo goed getraind was op de schijnpatiënten op het einde van de lijst dat het de gewenste gewichten in het begin van de lijst al het meest vergeten was.

---

<sup>9</sup>Ter informatie: de slices werden grotendeels geordend volgens hun slicehoogte. Echter met de eigenschap dat bijvoorbeeld slicehoogte 12.0 voor 120.0 komt. Ze werden als het ware “alfabetisch” geordend op slicehoogte. Dit gecombineerd met het feit dat er geen oneven slicehoogtes waren doordat de snededikte van een CT-scan 2 mm bedroeg, zorgde voor een opeenvolging van groepen van 10 anatomisch opeenvolgende slices met daartussen een “random” slice.

<sup>10</sup>De testgroep is de groep die compleet apart wordt gehouden van het trainproces van het model. Hier wordt achteraf het model op geëvalueerd. Op de validatieset wordt ook niet getraind. Hier wordt na bijvoorbeeld elke epoch de loss en score op berekend. Hiermee houden we het vermogen van het model om te kunnen generaliseren in het oog. Indien de validatieloss weer zou beginnen stijgen betekent dit dat het model aan het overtrainen is en dan stoppen we best met trainen.

## 2.2 De lossfunctie

Een eerste set van patiënten leverde, alle foto's samen, 65,280,295 pixels. Waarvan er 779,993 pixels of 1.19 procent deel uitmaakten van een orgaan. Een manier om 98.81 procent accuraatheid te hebben zou dus zijn om elke pixel het label “geen orgaan” te geven. Dit is niks nieuws aangezien neurale netwerken intrinsiek uitgaan van gebalanceerde data, wat hier niet het geval is. Er moeten dus stappen genomen worden om het klasse onevenwicht in rekening te brengen. Anders zal het neurale netwerk ons effectief een model leveren dat alles labelt als “geen orgaan”. Een neurale netwerk wil er namelijk vooral voor zorgen dat de mens gelooft dat het succesvol is in zijn taak. Klasseonevenwicht<sup>11</sup> kan op verschillende manieren opgelost worden. Hieronder bespreken we twee veelgebruikte methoden. Enerzijds het kiezen van een lossfunctie die accuraat ons doel weerspiegelt, namelijk het correct tekenen van oppervlakten. Anderzijds het harder straffen van het mislabelen van de minderheidsklasse, zijnde de organen.

### 2.2.1 Dixeloss

Een mogelijke lossfunctie is de “Dixeloss”. De Dicescore (zie sectie 2.4.1) op zichzelf is een maateenheid om te beschrijven hoe goed twee oppervlakten overeen komen. Een Dixeloss is dan makkelijk te definiëren als “hoe slecht twee oppervlakten overlappen”. De Dicescore (=DSC) is gedefinieerd als:

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|}$$

Waarbij  $|X|$  en  $|Y|$  het aantal elementen (pixels) is in elke set (contour) en  $|X \cap Y|$  het aantal elementen (pixels) die in beide sets (contouren) zitten. Dit levert dus een score van 0 tot en met 1. Een loss definiëren zou dus mogelijk zijn door 1 - de Dicescore te nemen. De exacte formule voor de Dixeloss (=DSL) voor één categorie ziet er als volgt uit:

$$DSL = 1 - \frac{2p\hat{p}}{p + \hat{p} + \epsilon}$$

Hierbij is  $p$  het al dan niet geactiveerd zijn van een pixel (0/1) en  $\hat{p}$  de kans dat het model denkt dat het dat orgaan is. De noemer heeft een  $+\epsilon$  gekregen om te vermijden dat we kunnen delen door nul. Dezelfde constante term zou toegevoegd kunnen worden in de teller om te garanderen dat de score vanaf 0 tot en met 1 loopt. Echter is deze  $\epsilon$  klein genoeg om binnen de accuraatheid dergelijke normalisatie te garanderen.

Indien we meerdere organen tegelijk willen classificeren moeten we een uitgemiddelde Dixeloss gebruiken [14]. Daarbij wordt de DSL voor elk orgaan apart berekend en vervolgens uitgemiddeld over het aantal organen.

---

<sup>11</sup>In dit geval het onevenwicht tussen het aantal pixels dat deel is van een orgaan en het aantal pixels dat niet deel is van een orgaan.

### 2.2.2 Gewogen binaire crossentropie loss

Binaire crossentropie loss (=BCEL) wordt gebruikt wanneer men twee labels heeft. In dat opzicht wordt het hier gebruikt wanneer we willen trainen op 1 orgaan (of tumor); het andere label is dan “niet dit orgaan”. Binaire crossentropie levert aan elke pixel dan de kans “deze pixel is orgaan”. De formule hiervoor is als volgt:

$$BCEL = -[p \cdot \log(\hat{p}) + (1 - p) \cdot \log(1 - \hat{p})]$$

Hierbij is  $p$  het al dan niet geactiveerd zijn van een pixel (0/1) en  $\hat{p}$  de kans dat het model denkt dat het dat orgaan is. Daarnaast worden  $p$  en  $\hat{p}$  geclipt in een interval  $[\epsilon, 1-\epsilon]$  zodat de lossfunctie overal reëel en afleidbaar blijft.

Deze lossfunctie gaat echter intrinsiek uit van klasse-evenwicht. Daarom gebruikt men bij dit soort problemen gewichten die representatief zijn voor het klasseonevenwicht. Dit zorgt voor de volgende formule voor de gewogen binaire crossentropie loss (=GBCEL):

$$GBCEL = -[p \cdot \log(\hat{p}) \cdot w_1 + (1 - p) \cdot \log(1 - \hat{p}) \cdot w_0]$$

Deze gewichten werden bepaald door de eerste 29 patiënten als voldoende grote steekproef. Indien er bijvoorbeeld 200 meer “niet long” pixels waren dan “long” pixels, moet het gewicht voor long 200 maal groter zijn. Daarnaast werd elke set van gewichten genormeerd. Dit gaf de volgende set van gewichten:

Tabel 3: Gewichten voor gewogen binaire crossentropie loss gehaald uit de labels van de eerste 29 patiënten.

Orgaan	$w_1$	$w_0$
Hart	416/417	1/417
Slokdarm	6792/679	1/6793
GTV	316895/316896	1/316896
Longen	110/111	1/111
Luchtpijp	6514/6515	1/6515
Ruggenmerg	6130/6131	1/6131

Men zou de vraag kunnen stellen waarom deze lossfunctie te gebruiken indien er al een andere functie is die expliciet rekening houdt met klasseonevenwicht. Echter heeft dit type aanpak al eerder zijn nut bewezen en daarom wordt ook deze piste geskied [13].

### 2.2.3 Gewogen categoriale crossentropie loss

Indien we niet één orgaan, maar meerdere willen trainen, moeten we noodgedwongen overgaan naar categoriale crossentropie loss (=CCEL). Deze laat meer dan twee categorieën toe. De formule ziet er hierbij als volgt uit:

$$CCEL = - \sum_{i=1}^N [p_i \cdot \log(\hat{p}_i)] - [1 - \sum_{i=1}^N p_i] \cdot \log[1 - \sum_{i=1}^N \hat{p}_i]$$

Hierbij gaat N over alle organen die we meetraineren. In het geval van N=1 krijgen we inderdaad de formule in sectie 2.2.2.

Ook deze formule heeft last van een intrinsieke verwachting aan klasse-evenwicht. Daarom werken we met gewichten die de volgende formule geven voor de gewogen categoriale crossentropie loss (=WCCEL):

$$WCCEL = - \sum_{i=1}^N [p_i \cdot \log(\hat{p}_i) \cdot w_{o1}] - [1 - \sum_{i=1}^N p_i] \cdot \log[1 - \sum_{i=1}^N \hat{p}_i] \cdot w_a$$

Voorgaande formules lijken niet meteen op de klassieke (gewogen) categoriale crossentropie loss formules. Deze laatste hebben niet het onderdeel  $[1 - \sum_{i=1}^N p_i] \cdot \log[1 - \sum_{i=1}^N \hat{p}_i]$ . Dit deel is echter wel nodig om een contragewicht te creëren ten opzichte van het classificeren als “wel een orgaan”. Dit deel zorgt er voor dat pixels ook worden gewogen tegen valse positieven. Concreet berekenen we eerst de klasseongelijkheid van “een orgaan” versus “niet een orgaan”. Zo komen we aan de  $w_a$  gewichten. Vervolgens verdelen we het gewicht van “wel orgaan” dat overblijft tussen alle zes de “wel orgaan” klassen waarbij de klasseongelijkheid tussen de verschillende organen gebruikt wordt voor de verdeling. Daarbij verkregen we volgende tabel:

Tabel 4: Gewichten voor gewogen categoriale crossentropie loss gehaald uit de labels van de eerste 29 patiënten.

Categorie	i	$w_{oi}$
Hart	1	0.1/85
Slokdarm	2	1.7/85
GTV	3	79/85
Longen	4	0.03/85
Luchtpijp	5	1.63/85
Ruggenmerg	6	1.54/85
Categorie		$w_a$
Achtergrond		1/85

## 2.3 Pretrained encodergewichten van ImageNet

Zoals vermeld werd er gebruik gemaakt van voorgetrainde encodergewichten voor het neurale netwerk. ImageNet doet niet het werk voor jou. Het levert slechts al een semi-geoptimaliseerd startpunt in de vorm van encodergewichten getraind op meer dan 1.2 miljoen afbeeldingen. Waarom is deze methode zo succesvol in het versnellen van het trainen? Vergelijk het met twee personen die willen leren voetballen. Ze hebben beiden nul ervaring hierin, echter heeft persoon A algemeen nul sportervaring maar heeft persoon B 10 jaar aan hardlopen gedaan. Beide achtergronden zijn niet expliciet kerntaken van een goede voetballer worden, alsnog heeft persoon B al een streepje voor aangezien die persoon al veel uithoudingsvermogen, reactievermogen en dergelijke heeft.

Het is een heuse taak voor een neurale netwerk om de optimale set aan gewichten te vinden indien deze van nul vertrekt. Vergelijk het met het oplossen van een set van vergelijkingen. Indien men wilt oplossen voor drie onbekenden, heeft men drie onafhankelijke vergelijkingen nodig. Zo ook heeft men, om optimaal een grote set aan parameters te bepalen, een grote set aan data nodig. Algemeen: hoe meer data hoe beter. En daar komen de voorgetrainde encodergewichten van ImageNet van pas. Deze zijn al getraind op een ontzettend grote dataset.

In theorie klinkt dit mooi, maar werkt het ook in de praktijk? Volgens vorige onderzoeken alvast wel [5]. Toch werd de proef op de som genomen. De longen zijn veruit de makkelijkste organen om te trainen, dus deze zullen onze testsubjecten zijn. Er wordt van elke groep (train, test en validatie) van geshuffelde schijnpatiënten 1 patiënt genomen. Deze worden elk één keer gebruikt als train-, validatie- en testset.

- Configuratie 1: het gebruik van pretrained encoders met als backbone VGG16 waarbij de encodergewichten ook getraind kunnen worden.
- Configuratie 2: het gebruik van de pretrained encoders met als backbone VGG16 waarbij de encodergewichten bevroren blijven. Dit heeft als doel aan te tonen dat de al geleverde encodergewichten redelijk dicht bij de optimale encodergewichten voor ons probleem zullen liggen.
- Configuratie 3: het gebruik van een UNet model waarbij er niks voorgetraind is. Het model is te vinden op GitHub<sup>12</sup>.

Elke combinatie van configuratie en patiënten werd driemaal getest.

Voor alle modellen werd als finale activatiefunctie “sigmoid” gebruikt en als optimaliseerder “Adam” met een lineaire rectifier van 0.0001. De lossfunctie was de Dixeloss zoals beschreven in sectie 2.2.1 en de metriek was de Dicescore zoals beschreven in sectie 2.4.1. Dit om alles uiteraard zoveel mogelijk hetzelfde te houden. Een uniform overzicht van de experimenten is te vinden in tabel 5.

---

<sup>12</sup><https://github.com/zhixuhao/unet/blob/master/model.py>



Tabel 5: Overzicht van alle voorgestelde experimenten met hun variabelen. Elk experiment wordt driemaal uitgevoerd.

Modelconf.	Patiëntconf.	Activatiefunctie	Optimiser	Lossfunctie	Metriek
1	P1, P2, P3	Sigmoid	Adam(0.0001)	Diceloss	Dicescore
2	P1, P2, P3	Sigmoid	Adam(0.0001)	Diceloss	Dicescore
3	P1, P2, P3	Sigmoid	Adam(0.0001)	Diceloss	Dicescore
1	P3, P1, P2	Sigmoid	Adam(0.0001)	Diceloss	Dicescore
2	P3, P1, P2	Sigmoid	Adam(0.0001)	Diceloss	Dicescore
3	P3, P1, P2	Sigmoid	Adam(0.0001)	Diceloss	Dicescore
1	P2, P3, P1	Sigmoid	Adam(0.0001)	Diceloss	Dicescore
2	P2, P3, P1	Sigmoid	Adam(0.0001)	Diceloss	Dicescore
3	P2, P3, P1	Sigmoid	Adam(0.0001)	Diceloss	Dicescore

Een belangrijk verschil is nog het aantal trainbare parameters in het encoderdeel van het netwerk. Waar de VGG16 backbone 14,714,688 trainbare parameters heeft, heeft het encodergedeelte van het U-Net model in configuratie 3 er normaal gezien 4,684 224. Daarom werd het encodergedeelte in configuratie 3 omhoog geschaald naar 18,731,904 parameters. Dit door de (64, 128, 256, 512) configuratie omhoog te schalen naar (128, 256, 512, 1024). Dit geeft configuratie 3 een voordeel. Het zal echter duidelijk worden dat dit voordeel niet sterk genoeg is in vergelijking met de prestaties van de configuraties met de voorgetrainde encodergewichten. Uiteraard maakt dit de U-Net backbone niet volledig equivalent met de VGG16 backbone, over de algemene architectuur kunnen we niet veel wijzigen. Het is wel een belangrijke stap in het experiment zo equivalent mogelijk maken.

## 2.4 Metrieken

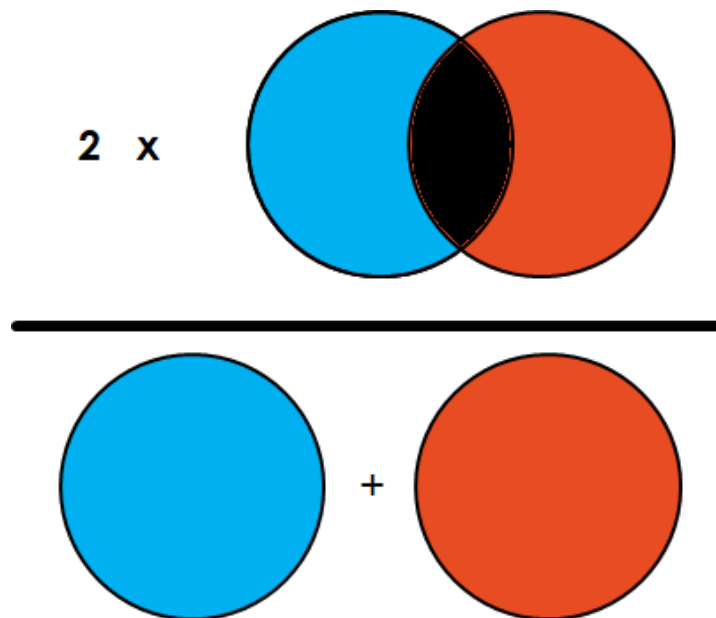
Na het trainen van het model moeten we kunnen oordelen of het model al dan niet geslaagd is in onze opzet. Daarvoor moeten we enkele metrieken definiëren. Welke metriek men gebruikt, is vaak afhankelijk van het doel van het experiment. Bij ons is dit correct oppervlakten intekenen. Daarom gebruiken we enkele klassieke oppervlaktemetrieken uit de wereld van machine learning. Ons finaal doel is echter niet zomaar zo perfect mogelijke intekeningen bekomen. Het is primair het appliceren van ons model in de radiotherapie. Daarom spenderen we extra aandacht aan enkele klinisch relevante metrieken.

### 2.4.1 Dicescore

Zoals eerder aangehaald is de Dicescore is gedefinieerd als:

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|}$$

Of visueel voorgesteld:



Figuur 9: Illustratie van de Dicescore [15].

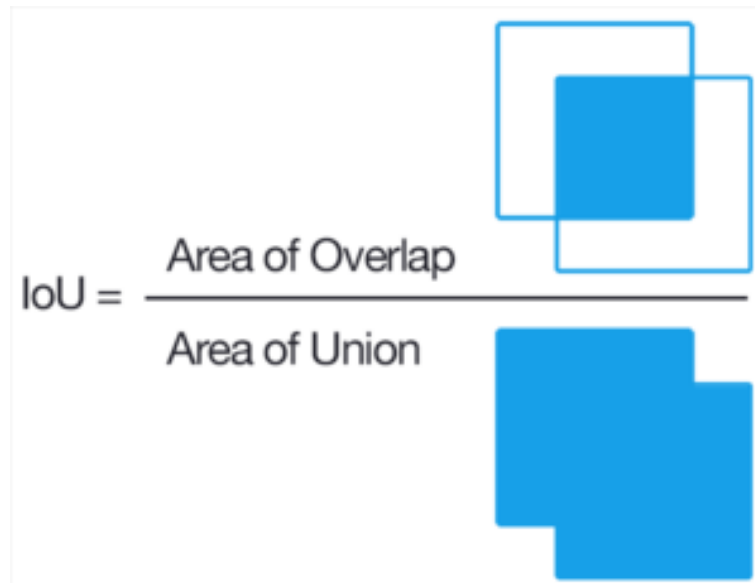
Deze score wordt gebruikt vanwege zijn directe link met de gebruikte Dixeloss functie. Naast deze gekende oppervlaktemetriek is er nog een andere bekende, de IoU (Intersection over Union) score. Deze metriek zien we vaak terugkeren in de literatuur, vandaar dat ze ook zal gebruikt worden.

### 2.4.2 IoU score

Deze score is als volgt gedefiniëerd:

$$IOU = \frac{|X \cap Y|}{|X| + |Y| - |X \cap Y|}$$

Vertaald betekent dit: gemeenschappelijk oppervlakte gedeeld door de unie van de 2 oppervlakten. Visueel voorgesteld door:



Figuur 10: Illustratie van de IoU Score [15].

Dit levert weer een 0 tot en met 1 score op. De IOU score is positief gecorreleerd met de Dicescore en zal gebruikt worden om onze resultaten te vergelijken met de resultaten uit de literatuur van sectie 1.4.3.

### 2.4.3 DVH en clinical goals

De DVH (Dosis Volume Histogram) is geen metriek die gebruikt wordt tijdens het trainen, maar een om achteraf te vergelijken bij de klinische evaluatie van de stralingsplannen. Dicescore en dergelijke zijn handig voor het trainen van een programma, maar vertellen weinig over de klinische gevolgen van de getekende organen. Het dosis volume histogram (DVH) van een orgaan is een 2-D weergave van de 3-D dosisverdeling binnen het orgaan. Voor elke dosiswaarde  $D$  wordt het relatieve volume  $V$  (in %) van het betreffende orgaan weergegeven dat een dosis krijgt groter dan of gelijk aan  $D$ . Klinische doelstellingen (clinical goals) worden vaak uitgedrukt in functie van dosis-volume parameters bijvoorbeeld: Longen:  $V_{20Gy} < 30\%$ . Dit wil zeggen dat het volume van de longen dat meer dan 20 Gy krijgt, niet meer dan 30% van het totale longvolume mag betreffen. Ruggenmerg:  $D_{2\%} < 50 \text{ Gy}$ . Dit wil zeggen dat slechts 2% van de voxels binnen het ruggenmerg meer dan 50 Gy mag krijgen. Men wil hierbij vooral de dosissen in normale, gezonde organen en weefsels in de buurt van de tumor(en) laag houden en die in tumor(en) hoog. Het doel is om de verschillende clinical goals te vergelijken tussen de ground truth en de voorspelde contouren. Voor de duidelijkheid: de ground truth zullen in dit geval de zelf opgestelde maskers zijn zoals beschreven in sectie 2.1.1. Dit om zo veel mogelijk fluctuaties tussen de pixelachtige organen en de meer fluente organen te vermijden.

## 2.5 Trainen

Graag voegen we een zekere disclaimer toe aan deze sectie. Wegens een bug die lang onopgemerkt is gebleven, zijn niet alle experimenten tot in het gewenste detail uitgevoerd kunnen worden. Er zijn afwegingen gemaakt tussen voldoende tests uitvoeren en voldoende resultaten hebben om de DHV en clinical goals toe te kunnen passen. Met dit in het achterhoofd lijsten we hier de uitgevoerde trainprocessen en hun parameters op.

### 2.5.1 Gewogen binaire crossentropie loss

Het belangrijkste om te weten is dat bij deze methode slechts één orgaan per model getraind kan worden. In dat opzicht zijn dit soort modellen minder interessant voor onze doelstelling van meervoudige segmentatie. Echter kan het in staat zijn om elk orgaan apart te segmenteren een indicator zijn van de haalbaarheid om meerdere organen tegelijk te segmenteren. Indien de gemakkelijke casus (één orgaan) niet werkt, kunnen we niet verwachten dat meerdere organen tegelijkertijd werken. Deze testen werden dus uitgevoerd met het oog op de “good practice” om opbouwend te werken.

Voor alle modellen werd opnieuw als finale activatiefunctie “sigmoid” gebruikt en als optimaliseerder “Adam” met een lineaire rectifier van 0.0001. De lossfunctie was de gewogen binaire crossentropie loss zoals beschreven in sectie 2.2.2 en de metriek was de IoU Score zoals beschreven in sectie 2.4.2 met een threshold van 0.5. Dit betekent dat een pixel als “orgaan” werd beschouwd indien het model minstens 50 procent zeker was.

Er werd gemodelleerd op twee categorieën patiëntengroepen. De eerste categorie bestaat uit drie schijnpatiënten waarbij elke groep (test, validatie en train) er één krijgt. Equivalent aan het systeem van 2.3. De tweede categorie bestaat uit de volledige 29 schijnpatiënten, waarbij de groep verdeeld werd in 19, 5 en 5 voor respectievelijk train, validatie en test zoals beschreven in 2.1.5. Welke schijnpatiënten voor welke groep gebruikt werd, was voor elk experiment hetzelfde. Voor alle organen werd hetzelfde principe gehanteerd. Er werd getest of er kon ge(over)traint worden op de configuratie met drie schijnpatiënten. Voor de longen werd er daarna ook getest op de configuratie met 29 schijnpatiënten. De andere categorieën kregen deze behandeling niet wegens tijdsnood.

Algemeen werden de longen en in zekere mate het hart het uitvoerigst getest. Helaas zijn de andere organen niet zo uitgebreid getest geweest wegens tijdsnood. Zij hebben hun tijd in zekere maten moeten inruilen voor experimenttijd voor sectie 2.5.2 en vooral voor sectie 2.5.3.

Voor alle categorieën werden enkel foto's gebruikt waar de categorie in voorkwam om de trainingstijd drastisch te verkorten. Voor de longen en het hart werden de volledige CT-beelden gebruikt. In de overige organen en de tumor werden enkel de centrale 256x256 pixels gebruikt zoals besproken in sectie 2.1.3. Dit aangezien er buiten deze regio geen organen te vinden zijn. Zo trachtten we de nog beschikbare tijd optimaal te benutten en voor de organen die overschieten een volwaardige test te doen op de configuratie met drie patiënten (één trainpatiënt). Uiteraard zullen de gewichten van sectie 2.2.2 voor deze aanpassingen opnieuw berekend worden. Voor de tumor werd echter nog een aparte test uitgevoerd waarbij alle foto's per patiënt werden gebruikt. De tumor heeft een fundamenteel ander karakter dan de kritische organen. Het lokaliseren van de tumor bepaalt namelijk sterk in welke richting er bestraald wordt. We willen hier dus extra aandacht geven aan hoge specificiteit.

### 2.5.2 Gewogen categoriale crossentropie loss

Na het testen van het trainbaar zijn van elk orgaan apart via crossentropie loss, is het de beurt aan het categoriale equivalent te testen. De lossfunctie die wij gebruiken is iets experimenteler dan de standaard categoriale crossentropie lossfunctie aangezien er ook gewichten zijn voor het “niet orgaan” zijn van elk orgaan. We hebben echter minstens één zo een contragewicht nodig aangezien het model anders getraind zal worden om elke pixel aan minstens één orgaan toe te schrijven. Elk orgaan krijgt een contragewicht zodat de verhouding tussen “orgaan x” en “niet orgaan x” behouden blijft.

Ook hier werd als finale activatiefunctie “sigmoid” gebruikt en als optimaliseerder “Adam” met een lineaire rectifier van 0.0001. De lossfunctie was de gewogen categoriale crossentropie loss zoals beschreven in sectie 2.2.2 en de metriek was de IoU Score zoals beschreven in sectie 2.4.2 met een van threshold 0.5. Dit betekent dat een pixel als “behorende tot deze categorie” werd beschouwd indien het model minstens 50 procent zeker was.

Er werd gemodelleerd op twee categorieën patiëntengroepen. De eerste categorie bestaat uit drie schijnpatiënten waarbij elke groep (test, validatie en train) er één krijgt. Equivalent aan het systeem van 2.3. De tweede categorie bestaat uit de volledige 29 schijnpatiënten, waarbij de groep verdeeld werd in 19, 5 en 5 voor respectievelijk train, validatie en test zoals beschreven in 2.1.5.

Enkel de centrale 256x256 pixels werden gebruikt zoals besproken in sectie 2.1.3. Uiteraard zullen de gewichten van sectie 2.2.3 voor deze aanpassingen opnieuw berekend worden.

### 2.5.3 Dixeloss

Voor de Dixeloss onderscheiden we weer de twee groepen van schijnpatiënten, met name de groep met drie en de groep met 29 patiënten. Opnieuw op dezelfde manieren onderverdeeld als in secties 2.5.1 en 2.5.2.

Eerst werd elk orgaan apart getraind om net zoals in sectie 2.5.1 de aparte trainbaarheden te evalueren. Hierbij werden weer enkel de foto's gebruikt waarop het orgaan geannoteerd was. Daarna werden alle organen tegelijkertijd getraind. Voor elk experiment werden enkel de centrale 256x256 pixels gebruikt.

Opnieuw werd als finale activatiefunctie “sigmoid” gebruikt en als optimaliseerder “Adam” met een lineaire rectifier van 0.0001. De lossfunctie was de Dixeloss zoals beschreven in sectie 2.2.1 en de metriek was de Dicescore zoals beschreven in sectie 2.4.1. Net zoals we bij de WCCEL een extra factor moesten invoeren zodat het model rekening kon houden met de achtergrond, moeten we bij het trainen op alle organen tegelijk ook een manier invoeren zodat het model “niet orgaan” kan voorspellen. Dit werd gedaan door *at runtime* een zevende “achtergrond” categorie toe te voegen. Deze was standaard *False* en werd *True* gemaakt indien alle andere categorieën *False* waren.

Daarnaast zal deze methode ook voor vijf patiënten (de vijf patiënten die de 5 geshuffelde schijnpatiënten opmaakten) getoetst worden aan de DHV en de clinical goals zoals beschreven in sectie 2.4.3.

### 3 Resultaten

Na het uitvoeren van een experiment is het tijd om uitvoerig de resultaten te bestuderen. Zijn we geslaagd in onze opzet? Hoe bepalen we of we geslaagd zijn? Waar is het misgegaan? Zijn onze resultaten in lijn met onze verwachten? Alle secties waarvan de experimenten waren opgelijst in sectie 2.5 zullen in deze sectie reeds uitvoerig besproken worden. Met die reden zal elke sectie een verwijzing hebben naar hun discussiesectie. Hierin zullen de voornaamste resultaten en bevindingen besproken worden.

#### 3.1 Pretrained ImageNet encoderweights

De resultaten van het experiment beschreven in sectie 2.3 zijn weergegeven in tabel 6.

Er werden drie schijnpatiënten gebruikt die afwisselend de train-, validatie- en testgroep waren. Waarbij elke mogelijke combinatie drie keer getest werd. We zullen deze drie patiënten “P1, P2 en P3” labelen.

De fout op de gemiddelde waarden is de steekproefstandaardafwijking.

Tabel 6: Dixeloss en Dicescore op de testset voor alle mogelijke combinaties van de drie schijnpatiënten en de drie configuraties, elk experiment werd driemaal herhaald.

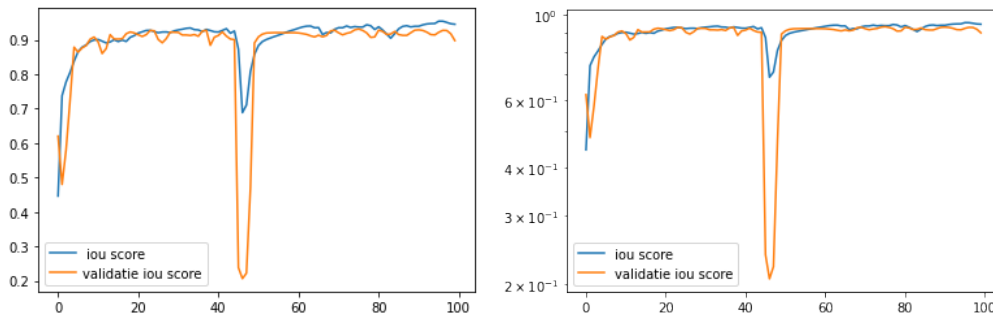
Modelconfiguratie	Patiëntconfiguratie	Gemiddelde test Dixeloss	Gemiddelde test Dicescore
1	P1, P2, P3	0.018 $\pm$ 0.027	0.961 $\pm$ 0.031
2	P1, P2, P3	0.029 $\pm$ 0.032	0.955 $\pm$ 0.033
3	P1, P2, P3	0.726 $\pm$ 0.017	0.236 $\pm$ 0.018
1	P3, P1, P2	0.064 $\pm$ 0.037	0.946 $\pm$ 0.032
2	P3, P1, P2	0.066 $\pm$ 0.005	0.959 $\pm$ 0.007
3	P3, P1, P2	0.823 $\pm$ 0.023	0.219 $\pm$ 0.021
1	P2, P3, P1	0.040 $\pm$ 0.063	0.965 $\pm$ 0.072
2	P2, P3, P1	0.042 $\pm$ 0.007	0.952 $\pm$ 0.012
3	P2, P3, P1	0.779 $\pm$ 0.035	0.257 $\pm$ 0.032

## 3.2 Gewogen binaire crossentropie loss

Voor elk orgaan worden verschillende kwalitatieve en kwantitatieve tests uitgevoerd. Deze sectie bevat een gedetailleerd overzicht van de resultaten. Een beknopte bespreking bevindt zich in sectie 4.2.

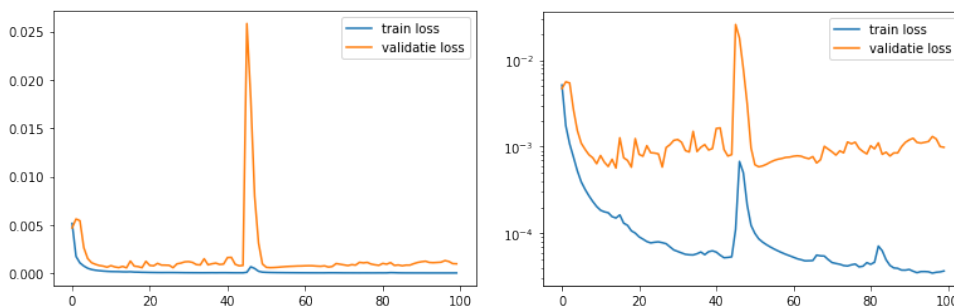
### 3.2.1 Longen

Voor de eerste test werd er voor 100 epochs getraind op de configuratie met drie schijnpatiënten. Gezien we op zoek waren naar een overtraind model, gebeurde het trainen slechts op één patiënt. We willen namelijk weten of het model trainbaar is en of het klaar is om getraind te worden op meer patiënten om zo een werkend model te verkrijgen. Het overtrainen zal zich minimaal uiten in het stagneren van de metrieken ten gevolge van de data-augmentatie en zich maximaal uiten in het uiteindelijk verslechteren van de validatiemetrieken indien het model alsnog in staat is de willekeurige variaties van de trainingsdata (ten gevolge van de data-augmentatie) te leren. Als eerste (visuele) indicator kijken we naar het verloop van de IoU score van de trainset en validatieset tijdens het trainen:



Figuur 11: Verloop van de IoU score op de trainset en de validatieset tijdens het trainen van de longen over een periode van 100 epochs. Lineaire schaal (links) en logschaal (rechts).

Op deze schaal kunnen we naar het einde toe al een lichte overtraining van de trainset ten opzichte van de validatieset zien. Het effect wordt pas visueel duidelijk als we naar het verloop van de loss waarde kijken:



Figuur 12: Verloop van de loss waarde op de trainset en de validatieset tijdens het trainen van de longen over een periode van 100 epochs. Lineaire schaal (links) en logschaal (rechts).

Op de figuur rechts is duidelijk te zien hoe vanaf epoch 50 de globale trend van de train loss nog steeds dalend (trainend is). Dit terwijl die van de validatie loss al globaal stijgend is. Dit hint naar overtrainen waarbij het model in staat is de fluctuaties ten gevolge van de data-augmentatie te overwinnen.

We zien een sterke piek in het verloop van de losswaarden. Dergelijke pieken kunnen verschillende verklaringen hebben. Kijkende naar onze situatie zijn er twee zeer logische verklaringen (die tegelijkertijd waar kunnen zijn). De eerste ligt bij de lossfunctie. Loglikelihood-losses moeten geclipped worden zodat evaluatie van de logfunctie in nul niet mogelijk is. Dit wordt bij ons gedaan. Echter kan het niet voldoende hard clippen van de waarden nog steeds voor evaluaties dicht bij nul zorgen. De tweede ligt bij de batchgrootte. Dit is geen exacte deler van het totaal aantal foto's van een epoch. Als gevolg hiervan kan de laatste batch van een epoch bestaan uit weinig foto's tot slechts één foto. Indien deze nog kleinere batch net heel anders is ten opzichte van de huidige gewichten van het model kan het een (onterecht) slechte score berekenen.

Er zijn ook kwantitatieve indicatoren voor het kunnen overtrainen van het model. De train IoU score bereikt een maximale waarde van 0.953 terwijl de test IoU score een waarde heeft van 0.881. Verder heeft de minimale train loss een waarde van 0.00003 terwijl de test loss een waarde heeft van 0.0006.

Als laatste kijken we naar de confusionmatrices. Deze vertelt ons voor de twee type labels (“long” en “geen long”) hoe goed het model deze correct kon labelen.

Tabel 7: Confusionmatrix van de testpatiënt na 100 epochs bij de longen.

Voorspelling/Ground Truth	Long (P = 1285274)	Niet Long(N = 30958438)
Long	True Positive (TP) = 1278272	False Positive (FP) = 121399
Niet Long	False Negative (FN) = 7002	True Negative (TN) = 30837039

Tabel 8: Confusionmatrix van de trainpatiënt na 100 epochs bij de longen.

Voorspelling/Ground Truth	Long (P = 985511)	Niet Long(N = 27326041)
Long	True Positive (TP) = 985508	False Positive (FP) = 110270
Niet Long	False Negative (FN) = 3	True Negative (TN) = 27215771

Indien we deze gegevens verwerken, bekommen we volgende kwantitatieve indicatoren:

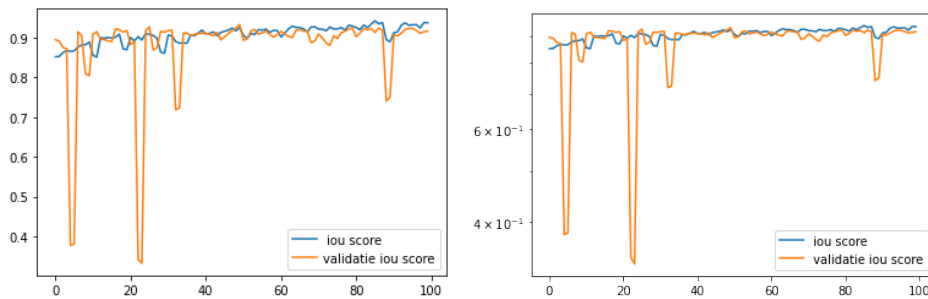
Tabel 9: True positive rate (TPR), true negative rate (TNR), false positive rate (FPR) en false negative rate (FNR) voor de test- en trainpatiënt na 100 epochs.

Patient	TPR = TP/P	TNR = TN/N	FPR = FP/N	FNR = FN/P
Test	0.950	0.996	0.004	0.050
Train	0.999	0.996	0.004	0.001



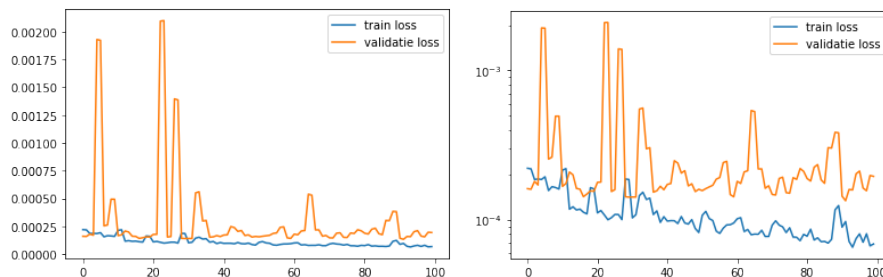
Hier zien we dat zowel de test- als de trainpatiënt competitieve TNR en FPR<sup>13</sup> hebben. Beiden zijn ze dus goed in het correct voorspellen van wat “niet long” is en gaan ze niet vlug iets “long” classificeren wat niet echt “long” is. De TPR en FNR rate van de testpatiënt is echter niet vergelijkend met die van de trainpatiënt. Dit betekent dat bij de testpatiënt de longen niet evengoed/volledig werden ingetekend als bij de trainpatiënt. Dit is een belangrijke metriek aangezien men in de radiotherapie liever wat te veel intekend dan te weinig. Men wil zo veel mogelijk garanderen dat er geen gezond weefsel wordt beschadigd. Algemeen kunnen we concluderen dat dit model overtraint kan worden en klaar is voor een grotere dataset.

Voor de tweede test werd er voor 100 epochs getraind op de configuratie met 29 schijnpatiënten (19 trainpatiënten). Dit was de logische uitbreiding na de conclusie dat het model te (over)trainen valt op 1 patiënt. Ook hier willen we vooral te weten komen of het model te (over)trainen valt op een grotere en gevarieerdere groep aan data. Als eerste (visuele) indicator kijken we naar het verloop van de IoU score van de trainset en validatieset tijdens het trainen:



Figuur 13: Verloop van de IoU score op de trainset en de validatieset tijdens het trainen van de longen over een periode van 100 epochs. Lineaire schaal (links) en logschaal (rechts).

Opnieuw kunnen we naar het einde toe net overtraining opmerken in de vorm van de train IoU score die blijft beteren terwijl de validatie IoU score stagneert. Een beter verschil zien we bij de losswaarden:



Figuur 14: Verloop van de loss waarde op de trainset en de validatieset tijdens het trainen van de longen over een periode van 100 epochs. Lineaire schaal (links) en logschaal (rechts).

Hier zien we dat de train loss nog steeds consistent aan het dalen is terwijl de validatie loss stagneert met nog een zekere fluctuatie ten gevolge van de data-augmentatie.

<sup>13</sup>We houden in het achterhoofd dat de TNR en de FPR elkaars complementaire zijn. Dit wil zeggen dat  $TNR + FNR = 1$ . Zo ook geldt dat  $TPR + FNR = 1$ .

We kijken verder naar de kwantitatieve indicatoren. De maximale train IoU score bedraagt 0.942 terwijl de test IoU score een waarde heeft van 0.916. Verder heeft de minimale train loss een waarde van 0.00006 terwijl de test loss een waarde heeft van 0.0006. De trainwaarden liggen in de trend van het vorige experiment, terwijl er een voorzichtige verbetering is in de testwaarden. Dit wijst naar het trainbaar zijn van het model op de grotere dataset.

We bekijken opnieuw de confusionmatrices binnen de testset en de trainset na 100 epochs. Om een betere vergelijking te kunnen, doen gebruiken we enkel de waarden van de patiënten die ook gebruikt werden in het vorige experiment.

Tabel 10: Confusionmatrix van de testpatiënt na 100 epochs bij de longen.

Voorspelling/Ground Truth	Long (P = 1285274)	Niet Long(N = 30958438)
Long	True Positive (TP) = 1280684	False Positive (FP) = 129191
Niet Long	False Negative (FN) = 4590	True Negative (TN) = 30829247

Tabel 11: Confusionmatrix van de trainpatiënt na 100 epochs bij de longen.

Voorspelling/Ground Truth	Long (P = 985511)	Niet Long(N = 27326041)
Long	True Positive (TP) = 985474	False Positive (FP) = 126234
Niet Long	False Negative (FN) = 37	True Negative (TN) = 27199807

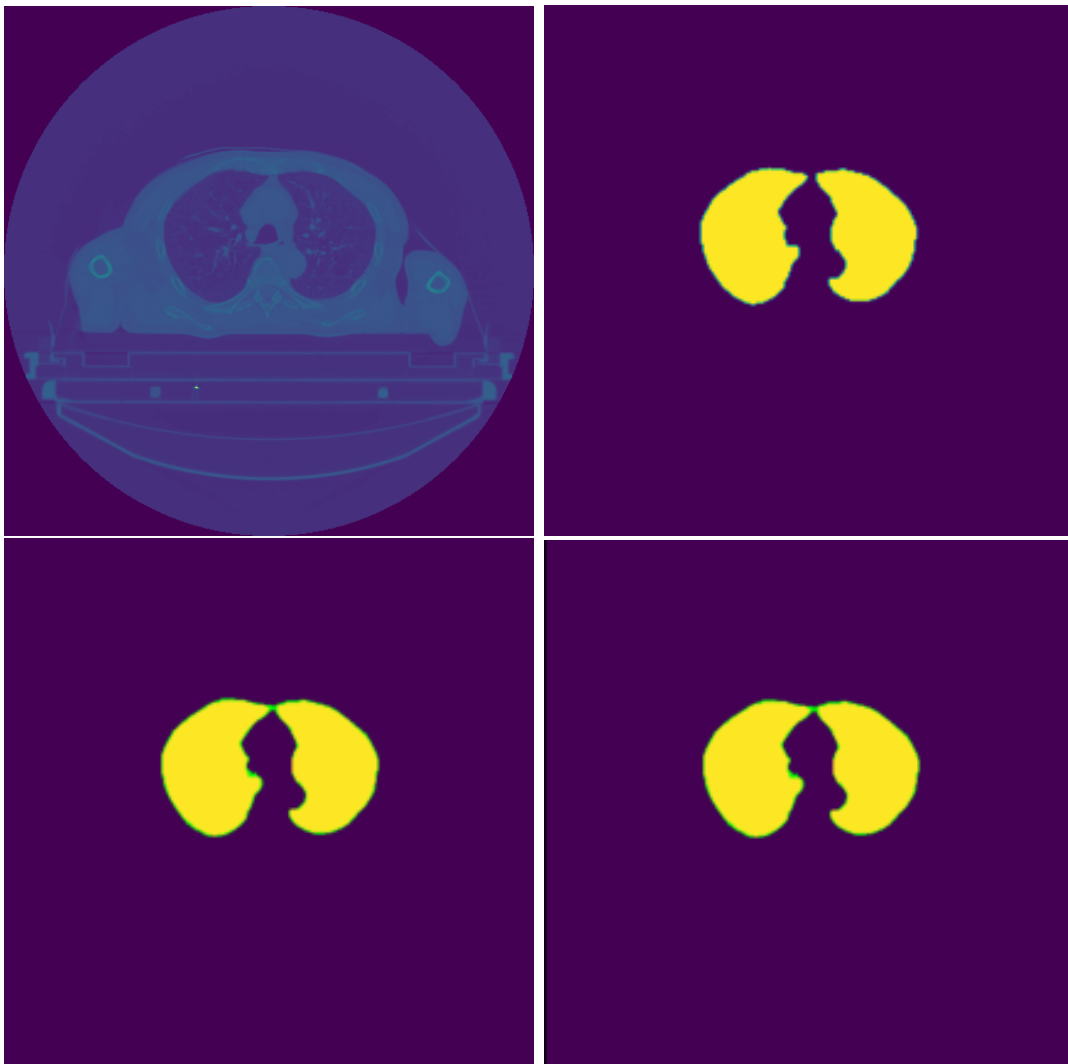
Indien we deze gegevens verwerken, bekommen we volgende kwantitatieve indicatoren:

Tabel 12: True positive rate (TPR), true negative rate (TNR), false positive rate (FPR) en false negative rate (FNR) voor de test- en trainpatiënt na 100 epochs.

Patient	TPR = TP/P	TNR = TN/N	FPR = FP/N	FNR = FN/P
Test	0.996	0.996	0.004	0.004
Train	0.99996	0.995	0.005	0.00004

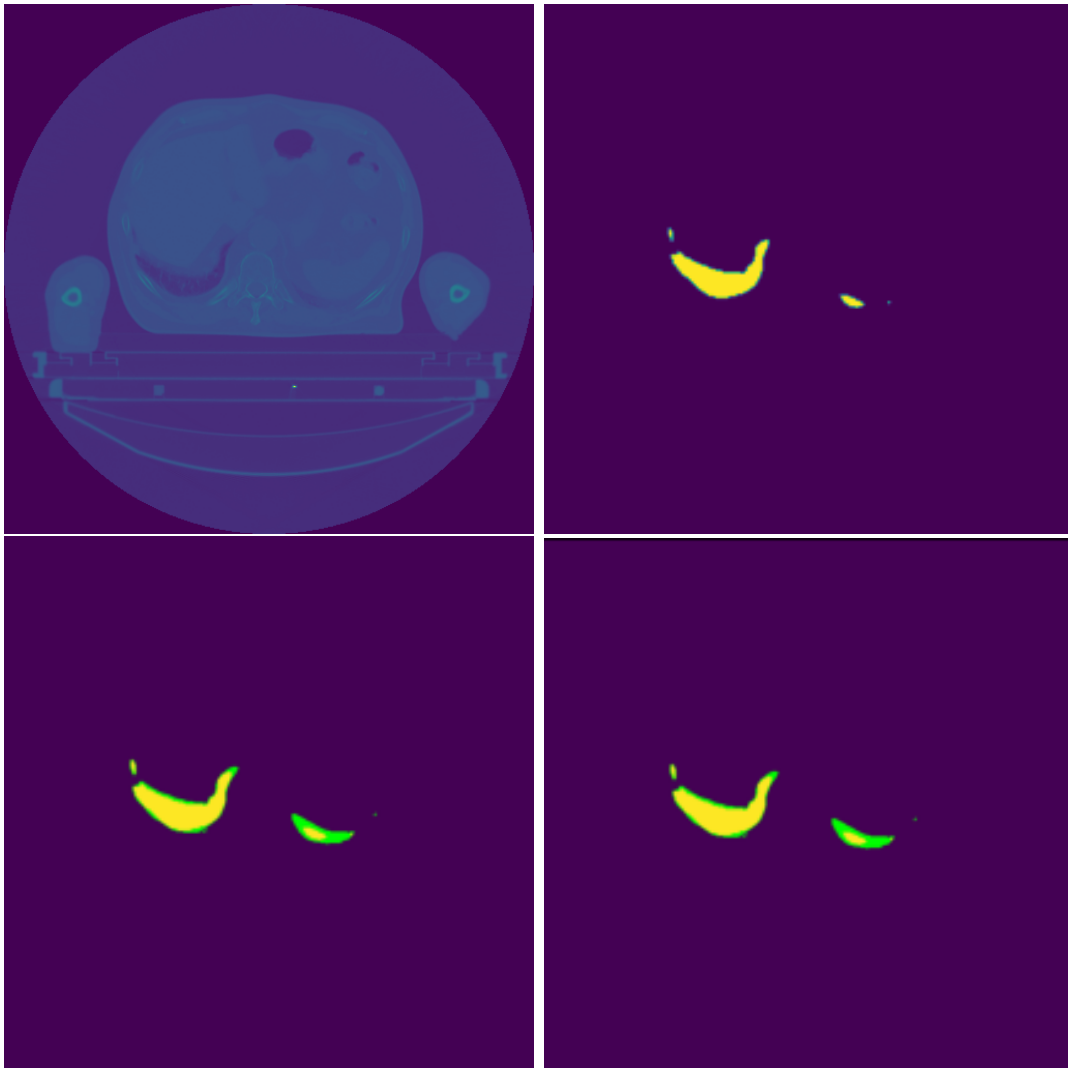
Hieruit kunnen we afleiden dat het model zeer goed is geworden in het voorspellen van de longen. De grotere dataset zorgde ervoor dat het model meer had om op te trainen. De data-augmentatie zorgde ervoor dat het model niet overtraine en generaliseerbaar bleef.

Als laatste geven we enkele visuele vergelijkingen tussen de ground truth segmentaties en de voorspellingen. Paars is TN, geel is TP, groen is FP en rood is FN. Beide vergelijkingen komen van de testpatiënt die ook in de confusionmatrix is gebruikt.



Figuur 15: CT-beeld (linksboven), ground truth longsegmentatie (rechtsboven), voorspelling experiment 1 (linksonder), voorspelling experiment 2 (rechtsonder).

We zien nog kleine fouten aan de randen van de voorspelde segmentaties. Dit zijn vooral valse positieven, wat minder erg is dan valse negatieven. Opmerkelijk is dat er amper verschil te zien is tussen de voorspellingen van het trainen op 1 patiënt (experiment 1) en het trainen op 19 patiënten (experiment 2).

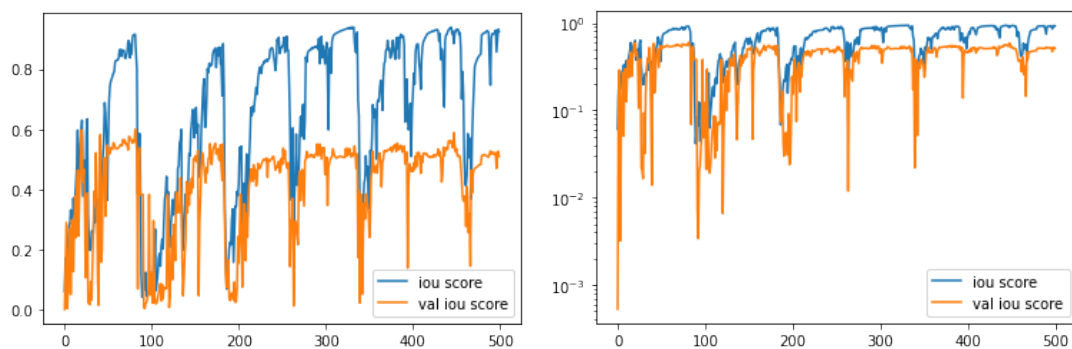


Figuur 16: CT-beeld (linksboven) ground truth longsegmentatie (rechtsboven) voorspelling experiment 1 (linksonder) voorspelling experiment 2 (rechtsonder).

Opnieuw zien we geen opmerkelijke verschillen tussen de voorspellingen van experiment 1 en experiment 2. We zien wel dat het model het moeilijker heeft met kleinere contouren. Het schat liever de longen groter in dan ze werkelijk zijn. Van valse negatieven is echter geen sprake wat wenselijk is want opnieuw geldt de redenering van liever valse positieven dan valse negatieven om de vitale organen zoveel mogelijk te beschermen tegen straling.

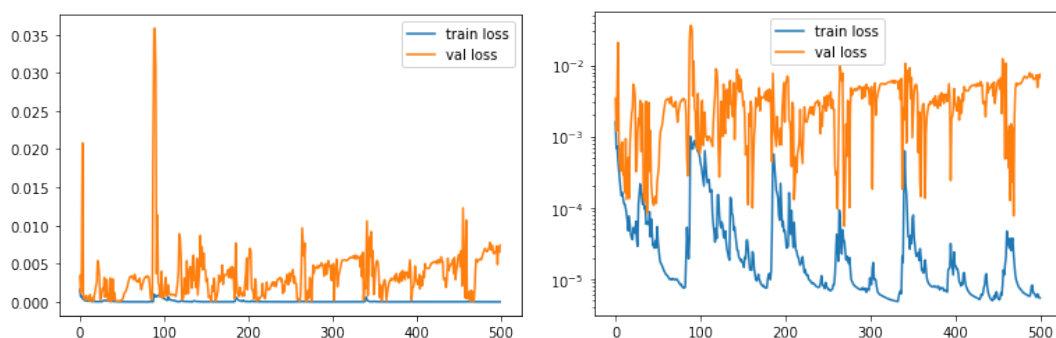
### 3.2.2 Hart

Voor het hart werd er weer gestart met de configuratie van drie schijnpatiënten. Echter werd direct voor 500 epochs getraind gezien de verwachte hogere moeilijkheidsgraad ten opzichte van de longen. Opnieuw waren we, gezien het trainen op slechts 1 patiënt, niet expliciet op zoek naar een “getraind” model, maar naar een overtraind model. We willen namelijk weten of het model trainbaar is en of het klaar is om getraind te worden op meer patiënten om zo een werkend model te verkrijgen. Het overtrainen zal zich minimaal uiten in het stagneren van de metrieken ten gevolge van de data-augmentatie en zich maximaal uiten in het uiteindelijk verslechteren van de validatiemetrieken indien het model alsnog in staat is de willekeurige variaties van de trainingsdata (ten gevolge van de data-augmentatie) te leren. Als eerste (visuele) indicator kijken we naar het verloop van de IoU score van de trainset en validatieset tijdens het trainen:



Figuur 17: Verloop van de IoU score op de trainset en de validatieset tijdens het trainen van het hart over een periode van 500 epochs. Lineaire schaal (links) en logschaal (rechts).

Opmerkelijk is dat de train IoU score consistent veel hoger ligt dan de validatie IoU score. Dit verschil was minder prominent bij de longen. Een mogelijke verklaring is grote verschillen tussen de foto's van de trainpatiënt en de validatiepatiënt. We zien in de rechtergrafiek een globale stagnatie vanaf epoch 200. Voor een duidelijker verschil kijken we naar de losswaarden:



Figuur 18: Verloop van de losswaarde op de trainset en de validatieset tijdens het trainen van het hart over een periode van 500 epochs. Lineaire schaal (links) en logschaal (rechts).

Bij de lineaire schaal zien we, zeker vanaf epoch 200, een globale stijgende trend in de validatie loss. Dit is ook te zien op de logschaal. Tevens zien we daar een duidelijke globale dalende trend voor de train loss zoals te verwachten is. Opmerkelijk is dat score en de loss voor zowel de trainpatiënt als de validatiepatiënt zeer volatiel is. Een mogelijke verklaring is weer enerzijds het gebruik van een loglikelihood-loss en anderzijds het niet perfect deelbaar zijn van de datasets in batches.

Aangezien de validatie loss en de validatie IoU score niet in de buurt komen van de waarden van de trainpatiënt kunnen we hetzelfde verwachten bij de loss en IoU score van de testpatiënt. De maximale IoU score van de trainpatiënt bedraagt 0.939 terwijl die van de testpatiënt 0.467 bedraagt. De minimale loss van de trainpatiënt bedraagt 0.000005 terwijl die van de trainpatiënt 0.014 bedraagt. Dit alles hint momenteel zeer sterk naar een sterke overtraining.

We kijken verder naar de confusionmatrices om te zien wat ze ons verder kunnen leren.

Tabel 13: Confusionmatrix van de testpatiënt na 500 epochs op het hart.

Voorspelling/Ground Truth	Hart (P = 241060)	Niet Hart(N = 10768988)
Hart	True Positive (TP) = 113668	False Positive (FP) = 12963
Niet Hart	False Negative (FN) = 127392	True Negative (TN) = 10756025

Tabel 14: Confusionmatrix van de trainpatiënt na 500 epochs op het hart.

Voorspelling/Ground Truth	Hart (P = 147537)	Niet Hart(N = 6406063)
Hart	True Positive (TP) = 146304	False Positive (FP) = 23412
Niet Hart	False Negative (FN) = 1233	True Negative (TN) = 6382651

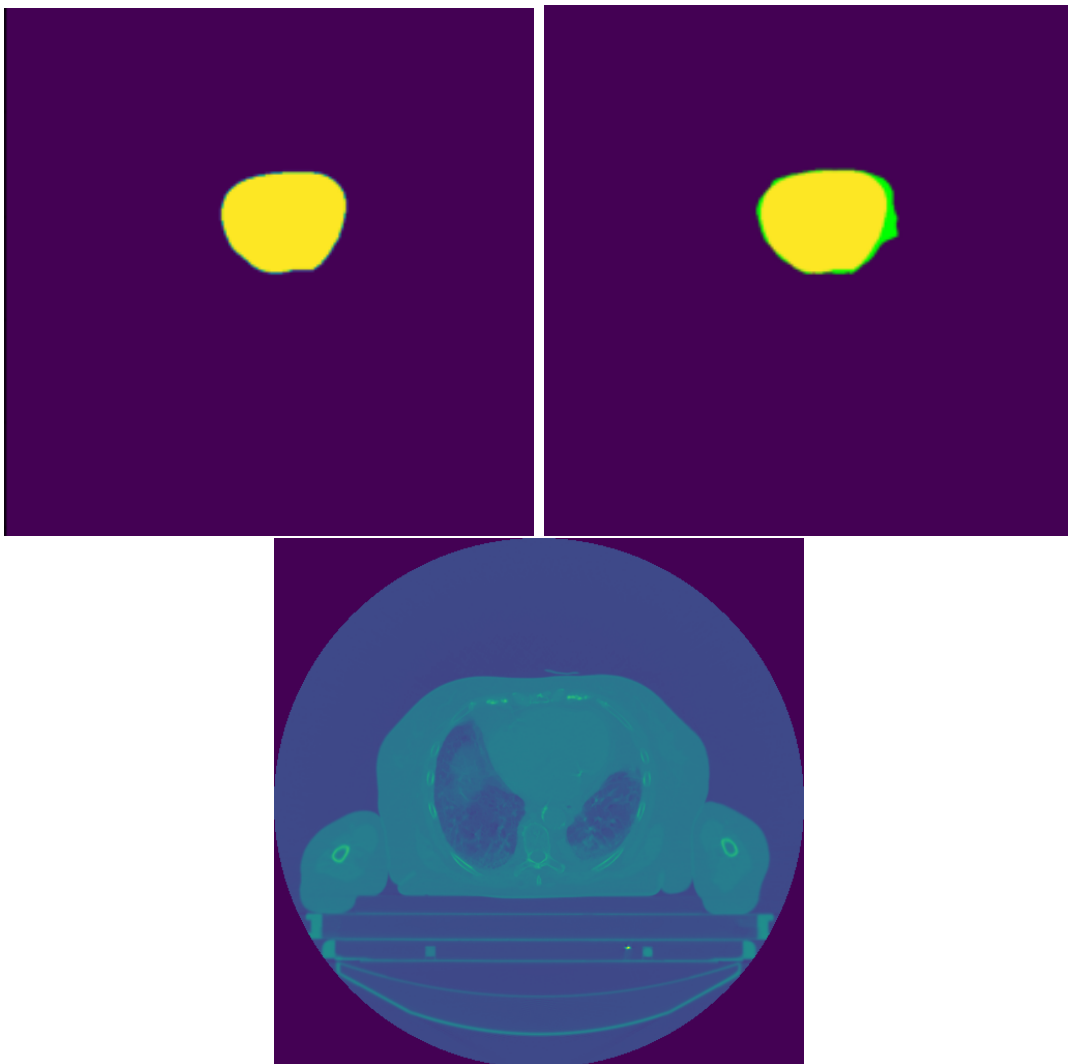
Indien we deze gegevens verwerken bekommen we volgende kwantitatieve indicatoren:

Tabel 15: True positive rate (TPR), true negative rate (TNR), false positive rate (FPR) en false negative rate (FNR) voor de test- en trainpatiënt na 500 epochs.

Patient	TPR = TP/P	TNR = TN/N	FPR = FP/N	FNR = FN/P
Test	0.472	0.999	0.001	0.528
Train	0.992	0.996	0.004	0.008

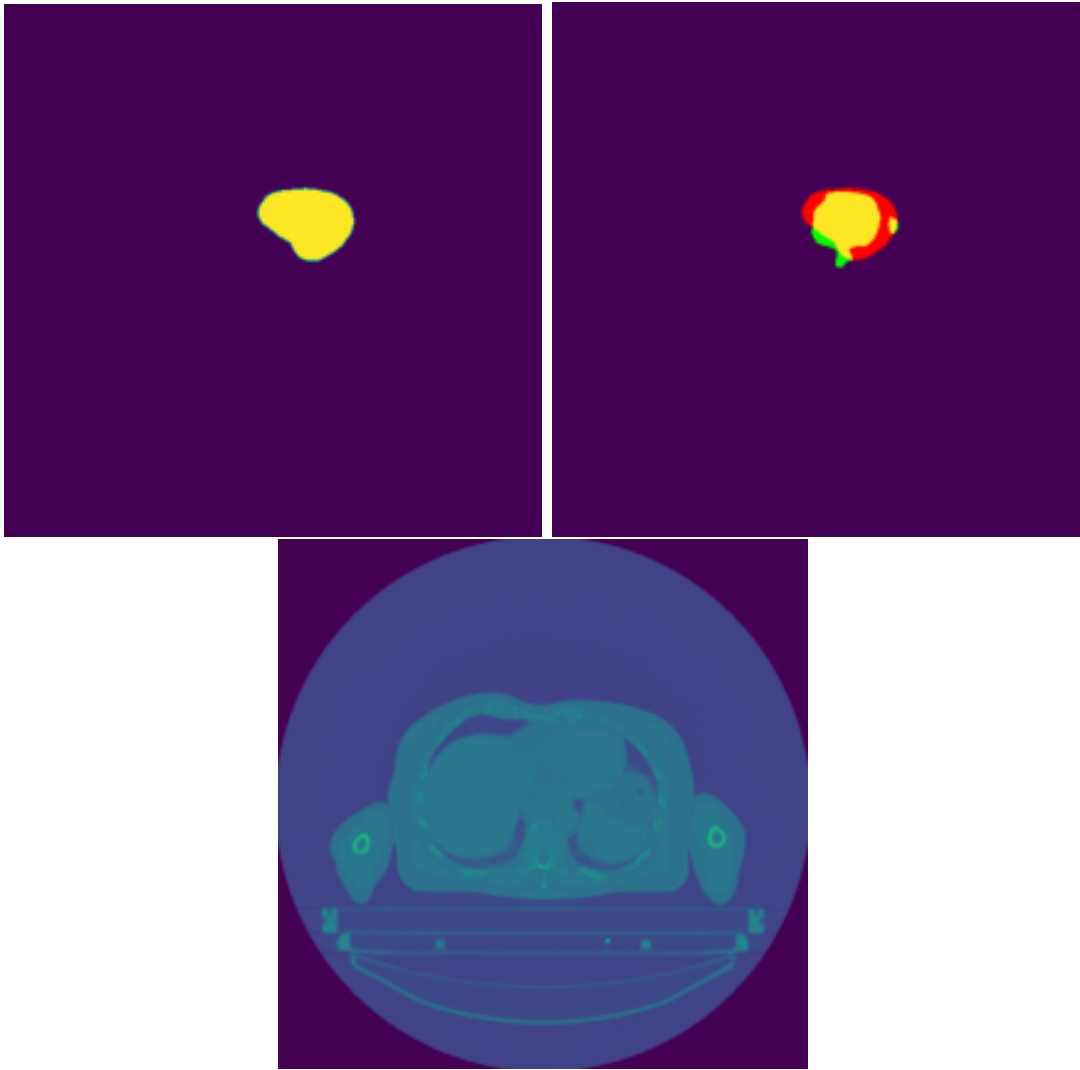
We zien weer competitieve TNR en FPR rates. Dit heeft echter weer vooral te maken met het feit dat de “niet hart”-klasse de dominante klasse is. De TPR en de FNR zijn echter verre van ideaal waarbij de FNR zelfs groter is dan de TPR.

Nog een laatste belangrijke controle is de visuele inspectie. Opnieuw vergelijken we de ground truth segmentaties en de voorspellingen. Paars is TN, geel is TP, groen is FP en rood is FN. De eerste set is afkomstig van de trainpatiënt en de tweede set van de testpatiënt.



Figuur 19: CT-beeld (onder) ground truth hartsegmentatie (linksboven) voorspelling experiment (rechtsboven) van de trainpatiënt.

We zien dat het model nog een te groot gebied voorspelt als “hart”. Dit is echter minder erg dan een te klein gebied voorspellen. Deze bevinding vinden we ook terug in de FPR ten opzichte van de FNR van de testpatiënt in tabel 15. We zien ook dat het hart minder eenvoudig te herkennen is op het CT-beeld in vergelijking met de longen.



Figuur 20: CT-beeld (onder) ground truth hartsegmentatie (linksboven) voorspelling experiment (rechtsboven) van de testpatiënt.

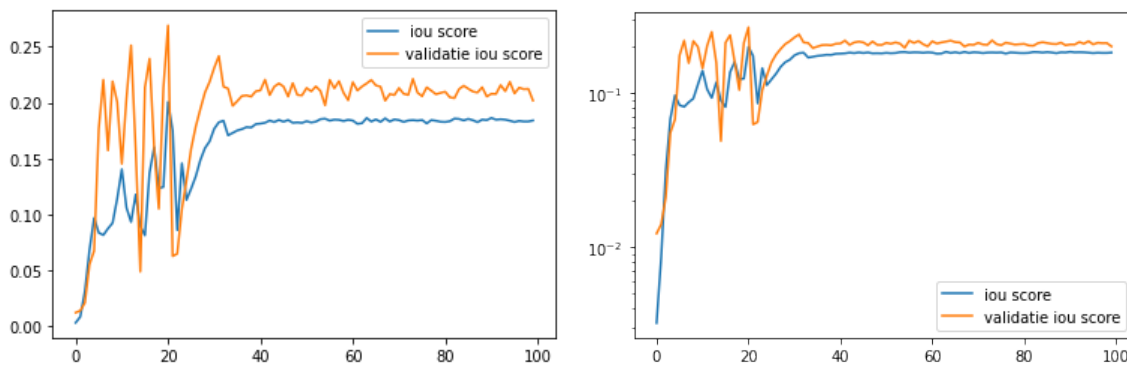
Dit resultaat is zeker niet wenselijk. Er wordt nog een te klein gebied voorspeld als “hart”. Valse negatieven zijn niet wenselijk aangezien men zeer goed moet weten waar de kritische organen zitten om deze niet te beschadigen tijdens de bestraling.

Zowel uit de trainingsgrafieken, de confusionmatrices als de visuele inspectie kunnen we concluderen dat het model in staat is om te overtrainen.



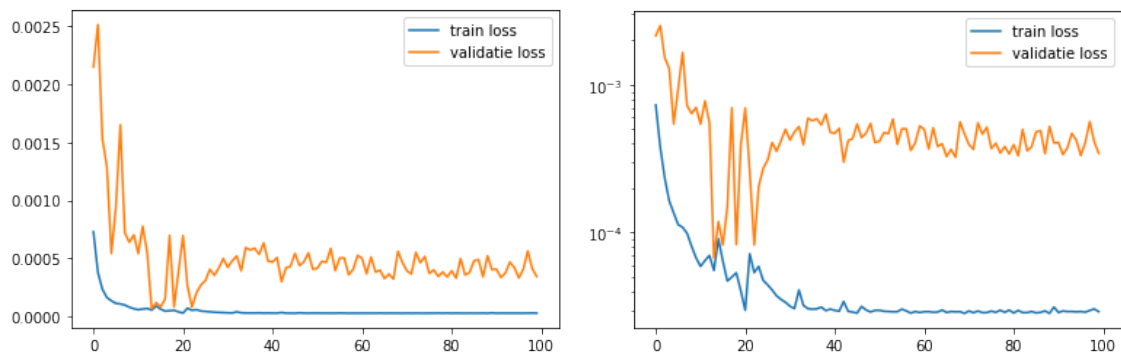
### 3.2.3 Slokdarm

Voor de slokdarm werd er weer gestart met de configuratie van drie schijnpatiënten en werd er getraind voor 100 epochs. Opnieuw waren we, gezien het trainen op slechts 1 patiënt, niet expliciet op zoek naar een “getraind” model, maar naar een overtraind model. We willen namelijk weten of het model trainbaar is en of het klaar is om getraind te worden op meer patiënten om zo een werkend model te verkrijgen. Het overtrainen zal zich minimaal uiten in het stagneren van de metrieken ten gevolge van de data-augmentatie en zich maximaal uiten in het uiteindelijk verslechteren van de validatiemetrieken indien het model alsnog in staat is de willekeurige variaties van de trainingsdata (ten gevolge van de data-augmentatie) te leren. Als eerste (visuele) indicator kijken we naar het verloop van de IoU score van de trainset en validatieset tijdens het trainen:



Figuur 21: Verloop van de IoU score op de trainset en de validatieset tijdens het trainen op de slokdarm over een periode van 100 epochs. Lineaire schaal (links) en logschaal (rechts).

We zien rond epoch 20 dat de validatie IoU score consistent hoger ligt dan de train IoU score. Dit is niet logisch aangezien er op de train patiënt getraind wordt. Vermoedelijk zitten er, tenminste voor dit orgaan, redelijk grote foutenvlaggen op deze datapunten. Vanaf epoch 40 zien we een stagnatie (met een zekere fluctuatie) van het model. Dit duidt op overtrainen waar de data-augmentatie ervoor zorgt dat de validatie IoU score niet terug begint te dalen. We kijken verder naar de lossgrafieken:



Figuur 22: Verloop van de losswaarde op de trainset en de validatieset tijdens het trainen op de slokdarm over een periode van 100 epochs. Lineaire schaal (links) en logschaal (rechts).

Ook hier zien we vanaf epoch 40 een stagnatie in zowel de train loss en in zekere mate een stagnatie in de validatie loss. Dit is weer typerend voor het (over)traint zijn van het model waar de data-augmentatie het overtrainen in toom houdt.

De maximale train IoU score bedroeg 0.201 terwijl de test IoU score 0.198 bedroeg. De minimale train loss bedroeg 0.00003 terwijl de test loss 0.0004 bedroeg. De scores zijn echter algemeen zeer laag.

Uit deze informatie zouden we al kunnen oordelen dat deze aanpak niet werkt voor het trainen op de slokdarm. Doch zullen we de confusionmatrices bekijken om te zien wat we er nog uit kunnen leren.

Tabel 16: Confusionmatrix van de testpatiënt na 100 epochs op de slokdarm.

Voorspelling/Ground Truth	Slokdarm (P = 13624)	Niet Slokdarm(N = 6998728)
Slokdarm	True Positive (TP) = 11450	False Positive (FP) = 45101
Niet Slokdarm	False Negative (FN) = 2174	True Negative (TN) = 6953627

Tabel 17: Confusionmatrix van de trainpatiënt na 100 epochs op de slokdarm.

Voorspelling/Ground Truth	Slokdarm (P = 10292)	Niet Slokdarm(N = 6739916)
Slokdarm	True Positive (TP) = 10292	False Positive (FP) = 46905
Niet Slokdarm	False Negative (FN) = 0	True Negative (TN) = 6693011

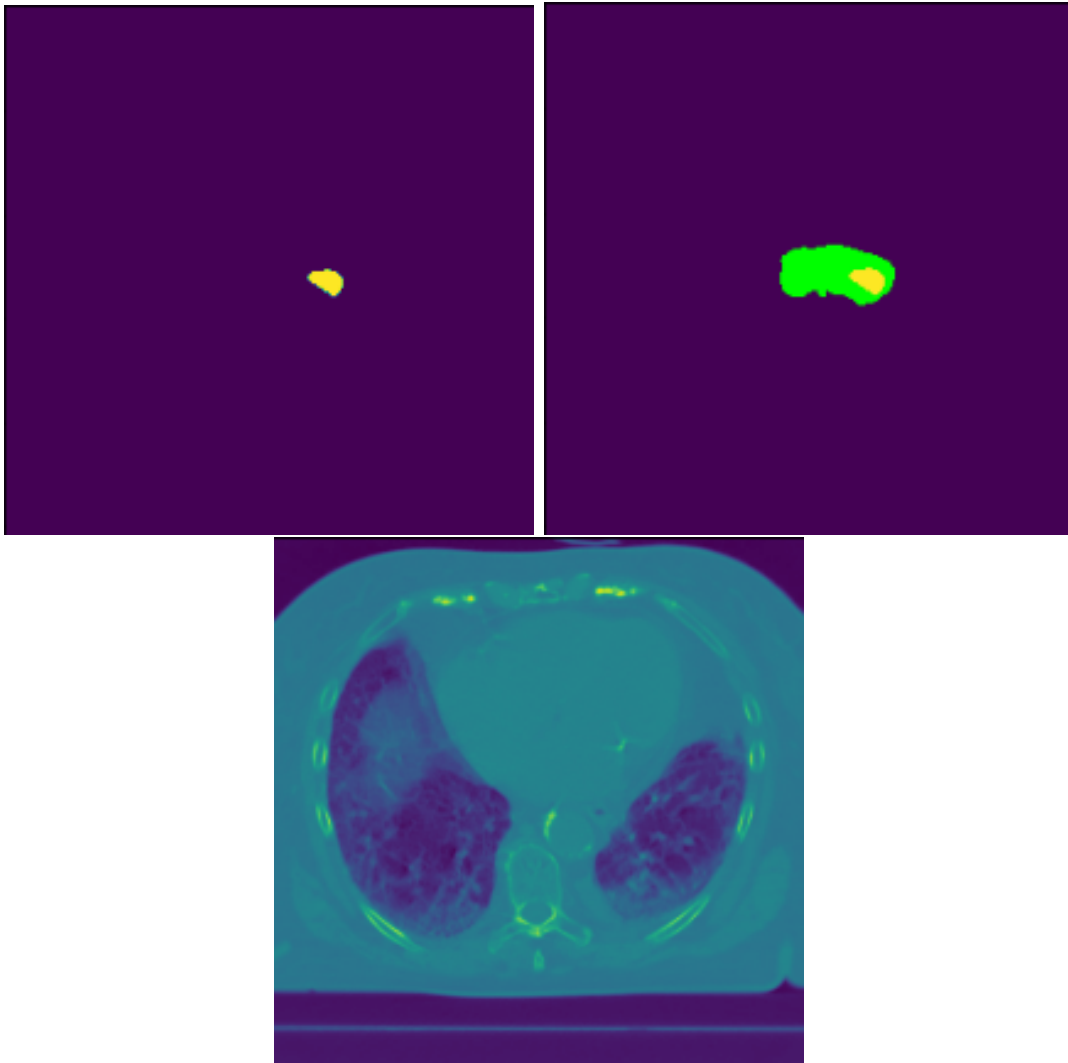
Indien we deze gegevens verwerken bekommen we volgende kwantitatieve indicatoren:

Tabel 18: True positive rate (TPR), true negative rate (TNR), false positive rate (FPR) en false negative rate (FNR) voor de test- en trainpatiënt na 100 epochs.

Patient	TPR = TP/P	TNR = TN/N	FPR = FP/N	FNR = FN/P
Test	0.840	0.994	0.006	0.160
Train	1.000	0.993	0.007	0.000

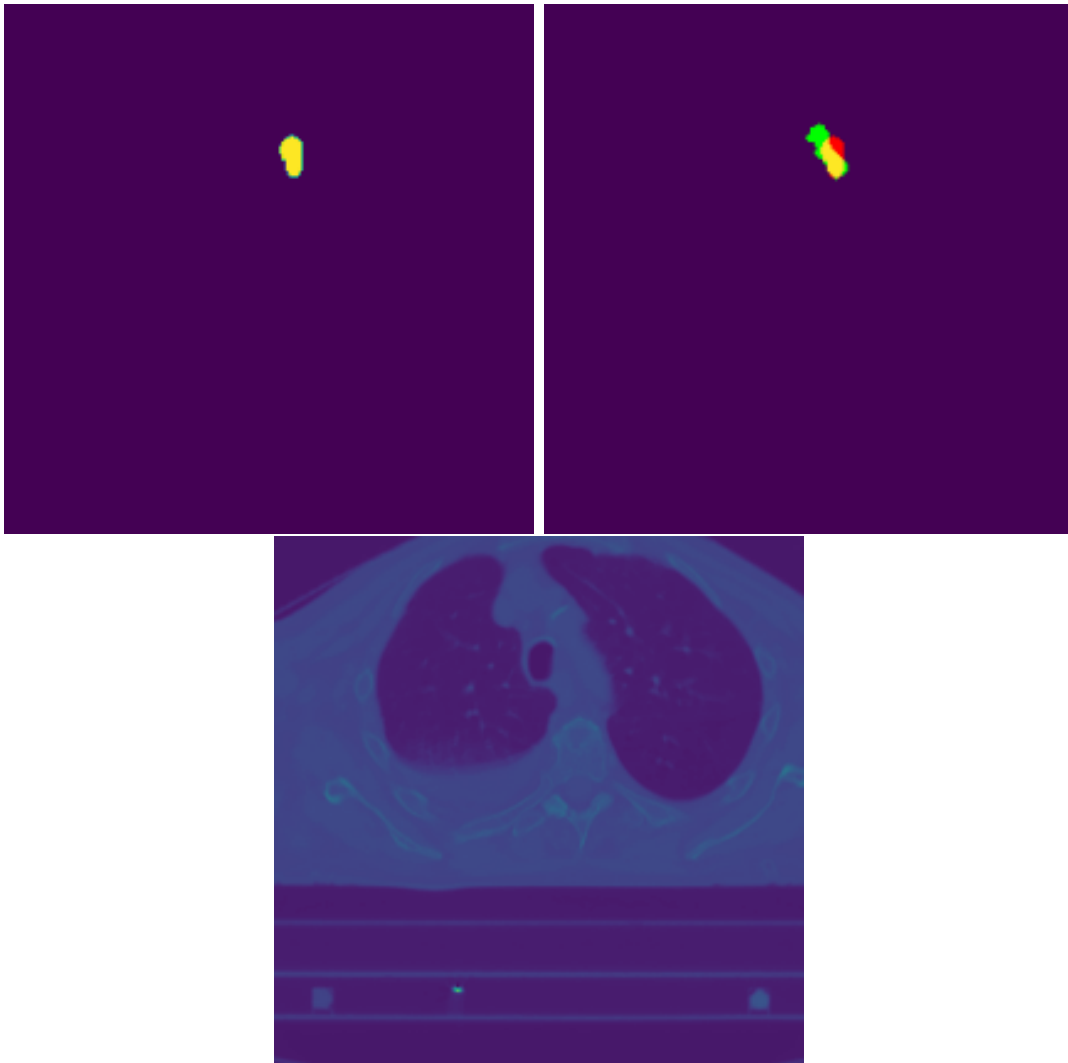
Hieruit blijkt dat de kwantitatieve indicatoren positiever uitdraaiden dan we hadden verwacht uit de kwalitatieve indicatoren. De trainpatiënt heeft een perfecte TPR wat betekent dat elke pixel die effectief een slokdarm is, ook zo werd teruggevonden. Ook de TPR van de testpatiënt is redelijk in orde. Slechter dan de testpatiënt wat dan weer hint naar overtraining.

Nog een laatste belangrijke controle is de visuele inspectie. Opnieuw vergelijken we de ground truth segmentaties en de voorspellingen. Paars is TN, geel is TP, groen is FP en rood is FN. De eerste set is afkomstig van de trainpatiënt en de tweede set van de testpatiënt.



Figuur 23: CT-beeld (onder) ground truth slokdarmsegmentatie (linksboven) voorspelling experiment (rechtsboven) van de trainpatiënt.

Zoals verwacht is er geen enkele FN. Er zijn echter een hoop valse positieven. Het orgaan wordt veel groter ingetekend dan het daadwerkelijk is. Dit verklaart de perfecte TPR maar de zeer lage IoU score. Het model was al redelijk in staat de generale positie van de slokdarm terug te vinden. Echter is verdere training en data nodig om verder te verfijnen. Zeker aangezien, zoals te zien is op het CT-beeld, de slokdarm redelijk opgaat in zijn omgeving.



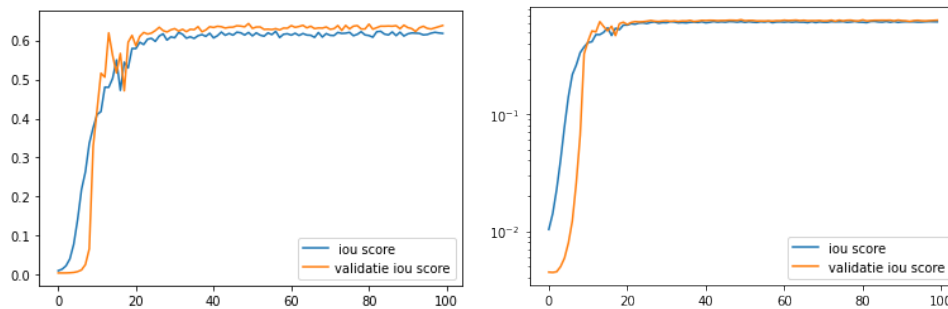
Figuur 24: CT-beeld (onder) ground truth slokdarmsegmentatie (linksboven) voorspelling experiment (rechtsboven) van de testpatiënt.

Zoals verwacht is er een relatief groot gebied met FP'en en FN'en. Ook hier gaat de opmerking op dat er vermoedelijk meer training en/of data nodig was.

Ondanks de initiële kwalitatieve indicator dat het model niet te overtrainen leek, hebben de kwantitatieve indicatoren en de beelden doen inzien dat er vermoedelijk simpelweg meer trainingstijd en data nodig is. Het resultaat bestempelen we daarom als onbepaald en verdere testen zullen nodig zijn.

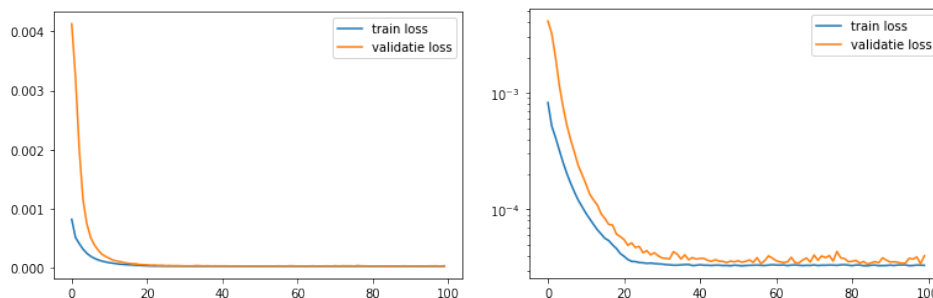
### 3.2.4 Luchtpijp

Voor de luchtpijp werd er weer gestart met de configuratie van drie schijnpatiënten en werd er getraind voor 100 epochs. Opnieuw waren we, gezien het trainen op slechts 1 patiënt, niet expliciet op zoek naar een “getraind” model maar naar een overtraind model. We willen namelijk weten of het model trainbaar is en of het klaar is om getraind te worden op meer patiënten om zo een werkend model te verkrijgen. Het overtrainen zal zich minimaal uiten in het stagneren van de metrieken ten gevolge van de data-augmentatie en zich maximaal uiten in het uiteindelijk verslechteren van de validatiemetrieken indien het model alsnog in staat is de willekeurige variaties van de trainingsdata (ten gevolge van de data-augmentatie) te leren. Als eerste (visuele) indicator kijken we naar het verloop van de IoU score van de trainset en validatieset tijdens het trainen:



Figuur 25: Verloop van de IoU score op de trainset en de validatieset tijdens het trainen van de luchtpijp over een periode van 100 epochs. Lineaire schaal (links) en logschaal (rechts).

We merken op dat het model al zeer vroeg in een optimum geraakt. Dit is niet zo vreemd aangezien de luchtpijp zeer gelijkend is met de longen. We weten al dat de longen zeer gemakkelijk te trainen zijn door het zeer verschillende medium en de aanzienlijke grootte. De trachea heeft hetzelfde medium en verschilt in dat opzicht vooral in grootte. We kijken of we deze bevindingen ook opmerken in de losswaarden:



Figuur 26: Verloop van de loss waarde op de trainset en de validatieset tijdens het trainen van de luchtpijp over een periode van 100 epochs. Lineaire schaal (links) en logschaal (rechts).

We zien al zeer vroeg dat zowel de testpatiënt als de validatiepatiënt stagneert mits kleine fluctuaties.

De maximale train IoU score bedroeg 0.623 terwijl de test IoU score 0.711 bedroeg. De minimale train loss bedroeg 0.00003 terwijl de test loss 0.0006 bedroeg. Opmerkelijk is dat de test IoU score beduidend groter is dan de maximale train IoU score. Dit zou wijzen op relatief grote foutenvlaggen op de IoU scores.

We kijken verder naar de confusionmatrices om meer duiding te krijgen.

Tabel 19: Confusionmatrix van de testpatiënt na 100 epochs op de luchtpijp.

Voorspelling/Ground Truth	Luchtpijp (P = 19981)	Niet Luchtpijp(N = 3781107)
Luchtpijp	True Positive (TP) = 18851	False Positive (FP) = 6360
Niet Luchtpijp	False Negative (FN) = 1130	True Negative (TN) = 3774747

Tabel 20: Confusionmatrix van de trainpatiënt na 100 epochs op de luchtpijp.

Voorspelling/Ground Truth	Luchtpijp (P = 13187)	Niet Luchtpijp(N = 3591293)
Luchtpijp	True Positive (TP) = 13187	False Positive (FP) = 8046
Niet Luchtpijp	False Negative (FN) = 0	True Negative (TN) = 3583247

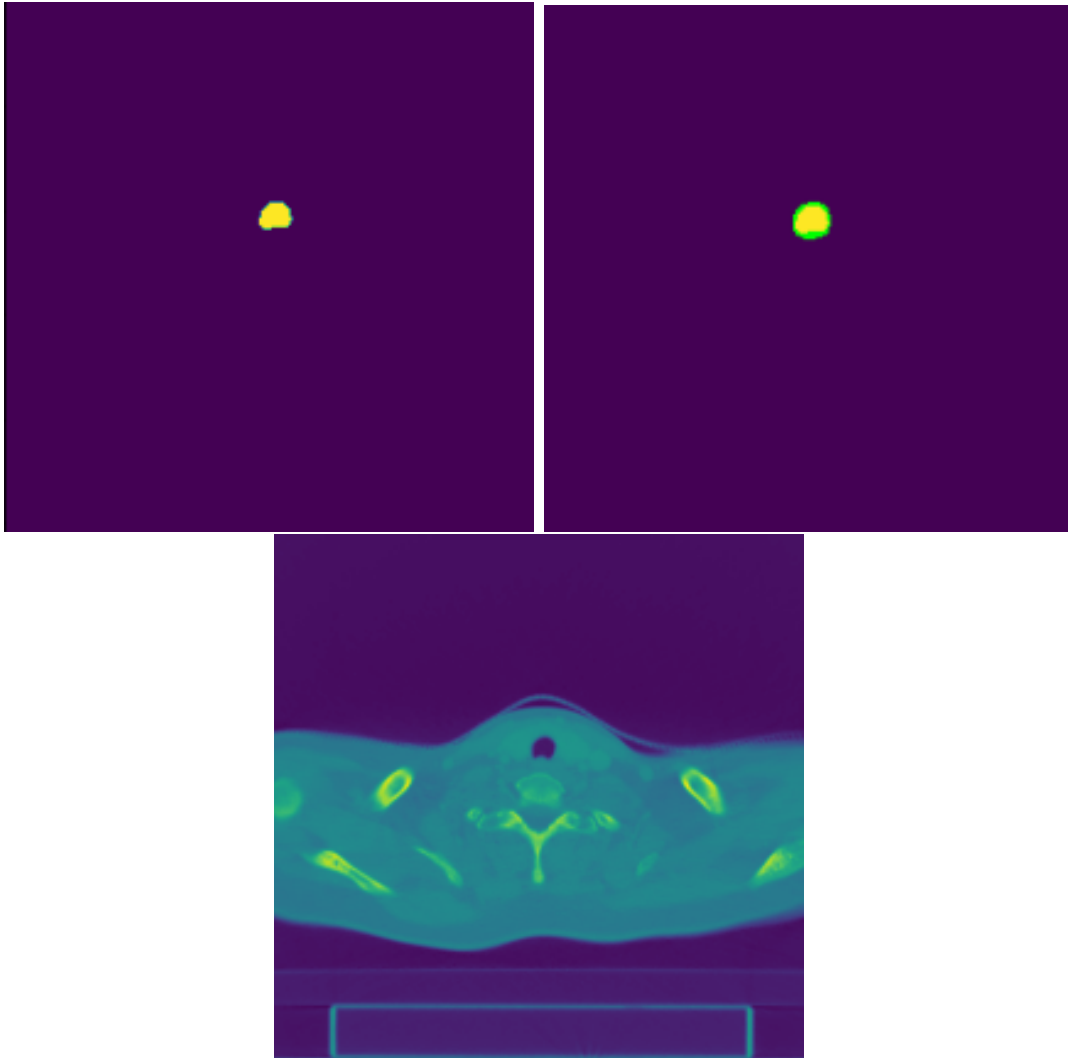
Indien we deze gegevens verwerken bekommen we volgende kwantitatieve indicatoren:

Tabel 21: True positive rate (TPR), true negative rate (TNR), false positive rate (FPR) en false negative rate (FNR) voor de test- en trainpatiënt na 100 epochs.

Patient	TPR = TP/P	TNR = TN/N	FPR = FP/N	FNR = FN/P
Test	0.943	0.998	0.002	0.057
Train	1.000	0.998	0.002	0.000

Opmerkelijk is dat we weer een perfecte TPR rate hebben bij de trainpatiënt zoals bij de slokdarm. Ook de TNR bij beide patiënten zijn competitief. Daarnaast is er een zeer hoge TPR voor de testpatiënt.

We vermoeden momenteel dat dit model zeker bruikbaar is. Het traint snel, heeft relatief hoge IoU scores en zeer goede TPR en TNR. Als laatste belangrijke indicator bekijken we weer de ground truth segmentaties en de voorspellingen. Paars is TN, geel is TP, groen is FP en rood is FN. De eerste set is afkomstig van de trainpatiënt en de tweede set van de testpatiënt.



Figuur 27: CT-beeld (onder) ground truth luchtpijpsegmentatie (linksboven) voorspelling experiment (rechtsboven) van de trainpatiënt.

Zoals we ook hebben gezien bij de slokdarm is er geen enkele FN. Echter is het model nu veel accurater in de FPR. Het model heeft bijna het orgaan verfijnd.



Figuur 28: CT-beeld (onder) ground truth luchtpijpsegmentatie (linksboven) voorspelling experiment (rechtsboven) van de testpatiënt.

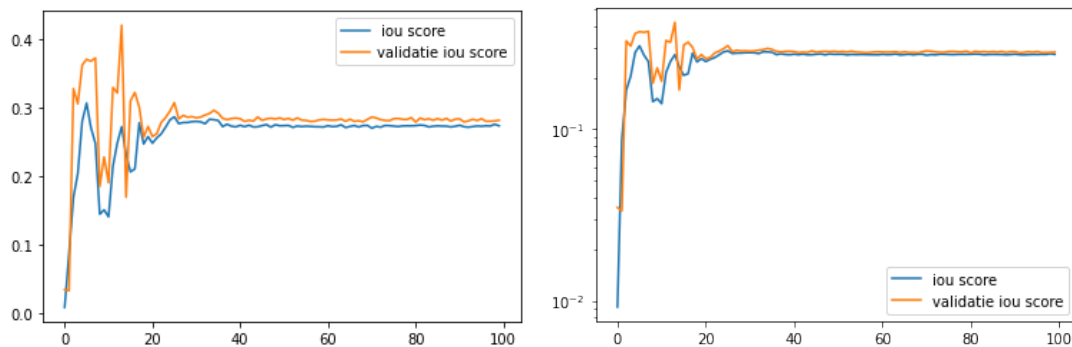
Ook bij de testpatiënt zien we een zeer goede segmentatie. We zien een dunne FN rand rond het orgaan. Gezien de grootte van deze rand is dit echter verwaarloosbaar ten opzichte van de precisie van de besturingsapparatuur.

Zowel uit de trainingsgrafieken, de confusionmatrices als de visuele inspectie kunnen we concluderen dat het model in staat is om te overtrainen en voor weinig data al zeer goede resultaten levert.



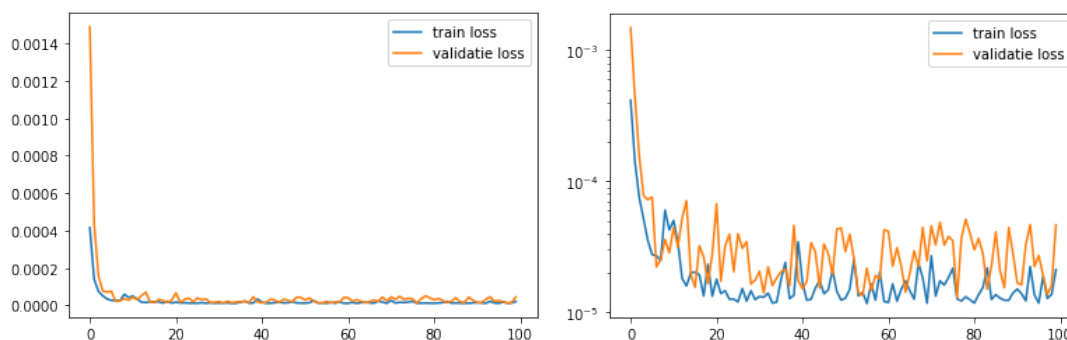
### 3.2.5 Ruggenmerg

Voor het ruggenmerg werd er weer gestart met de configuratie van drie schijnpatiënten en werd er getraind voor 100 epochs. Opnieuw waren we, gezien het trainen op slechts 1 patiënt, niet expliciet op zoek naar een “getraind” model maar naar een overtraind model. We willen namelijk weten of het model trainbaar is en of het klaar is om getraind te worden op meer patiënten om zo een werkend model te verkrijgen. Het overtrainen zal zich minimaal uiten in het stagneren van de metrieken ten gevolge van de data-augmentatie en zich maximaal uiten in het uiteindelijk verslechteren van de validatiemetrieken indien het model alsnog in staat is de willekeurige variaties van de trainingsdata (ten gevolge van de data-augmentatie) te leren. Als eerste (visuele) indicator kijken we naar het verloop van de IoU score van de trainset en validatieset tijdens het trainen:



Figuur 29: Verloop van de IoU score op de trainset en de validatieset tijdens het trainen van het ruggenmerg over een periode van 100 epochs. Lineaire schaal (links) en logschaal (rechts).

In het begin zijn er sterke fluctuaties en maxima in de IoU scores. Hier zitten vermoedelijk ook grote fouten op gezien het groter zijn van de validatie IoU score ten opzichte van de train IoU score. Rond epoch 30 lijkt zich dit te stagneren in een eerder lage IoU score zoals we hadden bij de slokdarm. Ook hier blijft de validatie IoU score consistent groter dan de IoU score. Vermoedelijk weer te verklaren door foutenvlaggen die we niet kunnen inschatten wegens tekort aan gereproduceerde data. We concluderen echter wel dat het model vroeg kan (over)trainen door de stagnaties van beide scores. We kijken verder naar de losswaarden:



Figuur 30: Verloop van de losswaarde op de trainset en de validatieset tijdens het trainen van het ruggenmerg over een periode van 100 epochs. Lineaire schaal (links) en logschaal (rechts).

Ook hier zien we een globale stagnatie van het model met zekere fluctuaties ten gevolge van de data-augmentatie. We kijken verder naar de kwantitatieve indicatoren.

De maximale train IoU score bedroeg 0.306 terwijl de test IoU score 0.345 bedroeg. De minimale train loss bedroeg 0.00001, net zoals de test loss. Opnieuw is de test IoU score beduidend groter dan de maximale train IoU score. De losswaarden zijn echter van hetzelfde kaliber. Iets wat we niet konden zeggen bij de vorige organen.

We kijken verder naar de confusionmatrices om meer duiding te krijgen.

Tabel 22: Confusionmatrix van de testpatiënt na 100 epochs op het ruggenmerg.

Voorspelling/Ground Truth	ruggenmerg (P = 20014)	Niet ruggenmerg(N = 14987730)
ruggenmerg	True Positive (TP) = 19849	False Positive (FP) = 40833
Niet ruggenmerg	False Negative (FN) = 165	True Negative (TN) = 14946897

Tabel 23: Confusionmatrix van de trainpatiënt na 100 epochs op het ruggenmerg.

Voorspelling/Ground Truth	ruggenmerg (P = 15602)	Niet ruggenmerg(N = 14598926)
ruggenmerg	True Positive (TP) = 15602	False Positive (FP) = 43507
Niet ruggenmerg	False Negative (FN) = 0	True Negative (TN) = 14555419

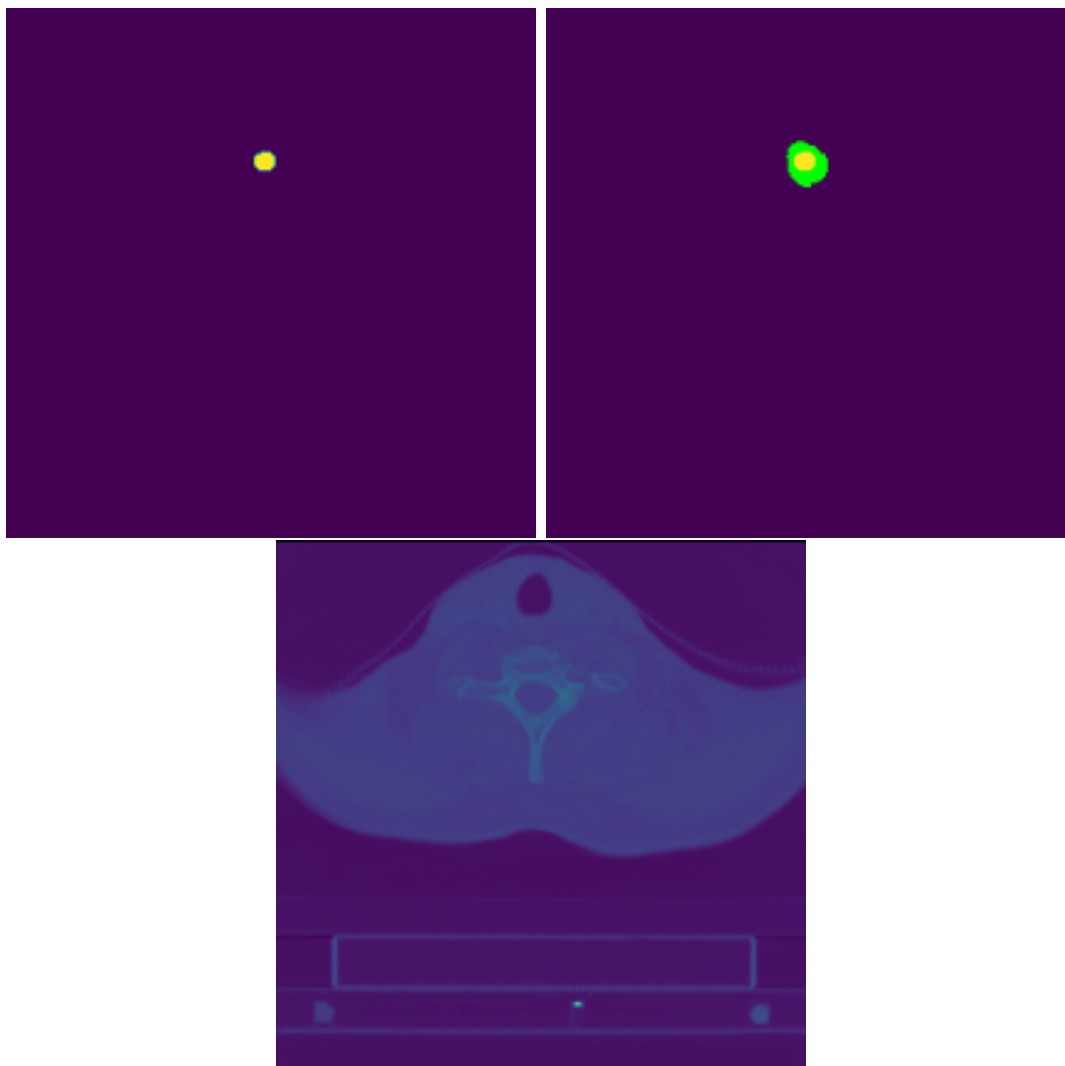
Indien we deze gegevens verwerken bekommen we volgende kwantitatieve indicatoren:

Tabel 24: True positive rate (TPR), true negative rate (TNR), false positive rate (FPR) en false negative rate (FNR) voor de test- en trainpatiënt na 100 epochs.

Patient	TPR = TP/P	TNR = TN/N	FPR = FP/N	FNR = FN/P
Test	0.992	0.997	0.003	0.008
Train	1.000	0.997	0.003	0.000

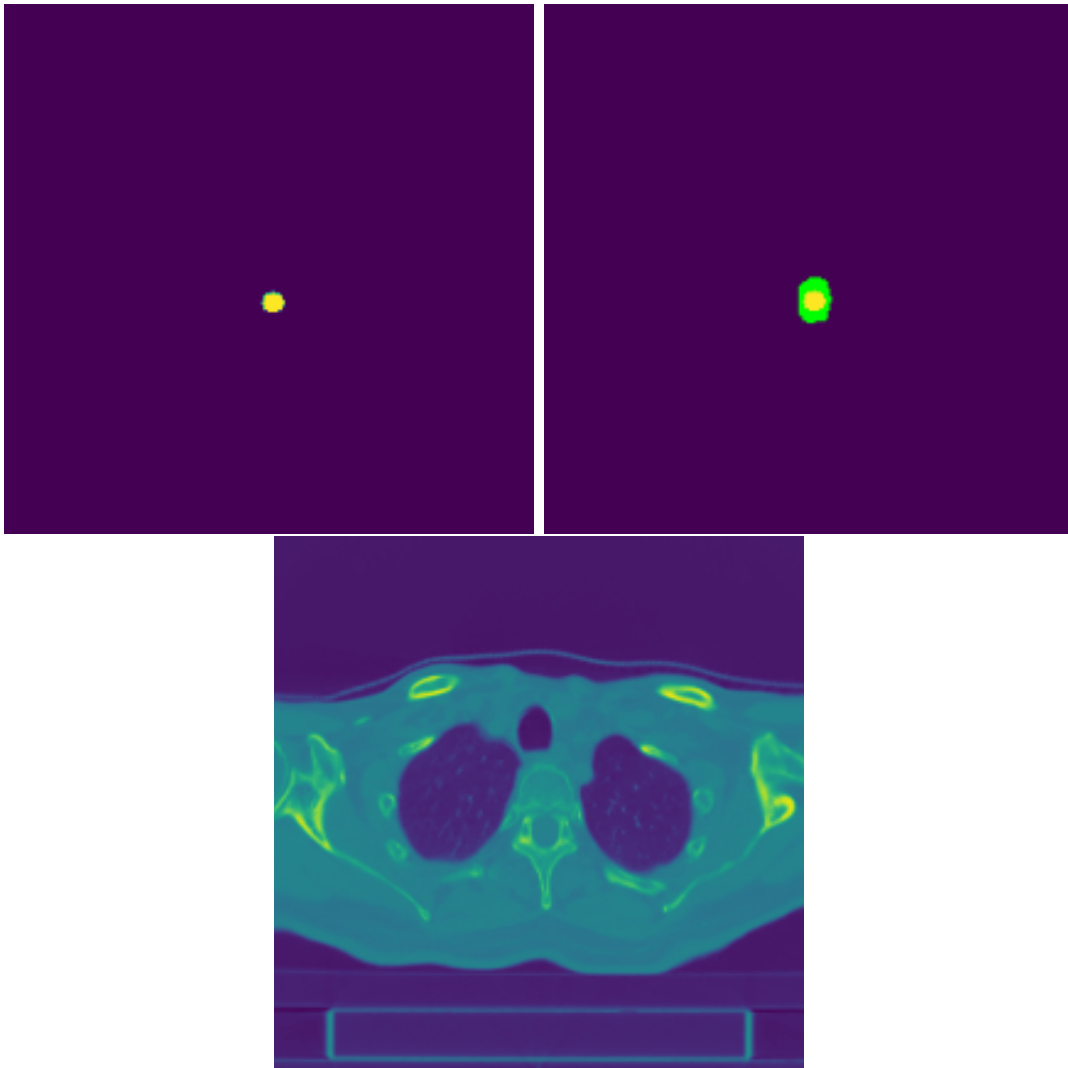
We merken hetzelfde op als bij de luchtpijp. Competitieve TNR en een vrij uitstekende TPR op de testset.

Aangezien de kwantitatieve indicatoren gelijkaardig zijn aan die van de luchtpijp verwachten we beelden van dezelfde aard. We zullen dit nu verifiëren. Paars is TN, geel is TP, groen is FP en rood is FN. De eerste set is afkomstig van de trainpatiënt en de tweede set van de testpatiënt.



Figuur 31: CT-beeld (onder) ground truth ruggenmergsegmentatie (linksboven) voorspelling experiment (rechtsboven) van de trainpatiënt.

Ondanks de TPR van de testpatiënt die meer in lijn lag van de luchtpijp, merken we eerder beelden op in lijn van de slokdarm. Het model heeft duidelijk de globale locatie van het orgaan gevonden. Echter duidt het nog een significant groot deel aan dat niet ruggenmerg is.



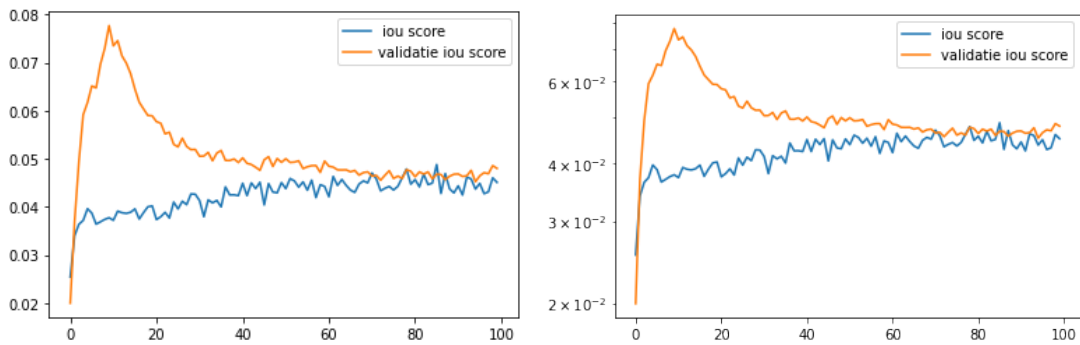
Figuur 32: CT-beeld (onder) ground truth ruggenmergsegmentatie (linksboven) voorspelling experiment (rechtsboven) van de testpatiënt.

Bij de trainpatiënt merken we dezelfde trend als bij de testpatiënt. We vermoeden weer dat er meer training en meer beelden nodig is om verder de locatie van het orgaan te specificeren.

Algemeen zouden we dit orgaan verder willen testen om te weten te komen of het orgaan verder te verfijnen is. De initiële resultaten lijken namelijk goed, maar er is nog redelijk wat uitlijning nodig. Het model lijkt in eerste instantie wel (over)trainbaar.

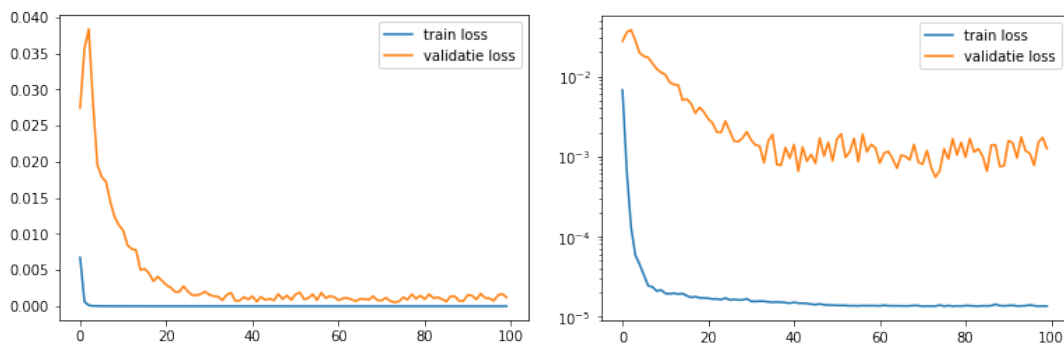
### 3.2.6 GTV

Voor het GTV werd er weer gestart met de configuratie van drie schijnpatiënten en werd er getraind voor 100 epochs. Net zoals de vorige organen bestond dit experiment enkel uit foto's waar effectief de categorie op te vinden was. Aangezien we voor het GTV ook zeer geïnteresseerd zijn in het correct kunnen zeggen wanneer iets geen tumor is, volgt er nog een ander experiment waar wel alle foto's per patiënt gebruikt werd. Voor het GTV hebben we geen expliciete verwachtingen gezien het compleet ander karakter dan de organen.



Figuur 33: Verloop van de IoU score op de trainset en de validatieset tijdens het trainen van het GTV over een periode van 100 epochs. Lineaire schaal (links) en logschaal (rechts).

We zien een consistent stijgende IoU score van de tumor wat duidt op een normaal trainproces. De validatie IoU score heeft een vreemde piek in het begin maar dit zal meer te wijten zijn aan statistische fluctuaties. Daarna daalt het gestaag om net op het einde weer lichtjes te beginnen stijgen. Voor meer informatie kijken we naar de losswaarden:



Figuur 34: Verloop van de losswaarde op de trainset en de validatieset tijdens het trainen van het GTV over een periode van 100 epochs. Lineaire schaal (links) en logschaal (rechts).

Op de lineaire schaal is er niet meteen iets duidelijk te zien. Op de logschaal zien we echter dat de train loss nog steeds globaal aan het dalen is. Dit staft ons vermoeden bij de IoU score dat het model nog steeds aan het trainen was op de trainpatiënt. De validatie loss is te grillig om met voldoende zekerheid een uitspraak te doen. Met dit in het achterhoofd kijken we verder naar meer kwantitatieve indicatoren.

De maximale train IoU score bedroeg 0.049 terwijl de test IoU score 0.013 bedroeg. De minimale train loss bedroeg 0.00001 terwijl de test loss 0.0002 bedroeg. Deze bevindingen wijzen wel sterk in het overtraint zijn van een model. We zijn echter niks met een overtraint model indien zelfs de trainset zeer slechte resultaten levert.

We kijken verder naar de confusionmatrices om meer inzicht te krijgen in de zeer lage IoU scores.

Tabel 25: Confusionmatrix van de testpatiënt na 100 epochs op het GTV.

Voorspelling/Ground Truth	GTV (P = 4768)	Niet GTV(N = 978272)
GTV	True Positive (TP) = 4768	False Positive (FP) = 319259
Niet GTV	False Negative (FN) = 0	True Negative (TN) = 659013

Tabel 26: Confusionmatrix van de trainpatiënt na 100 epochs op het GTV.

Voorspelling/Ground Truth	GTV (P = 26261)	Niet GTV(N = 1612139)
GTV	True Positive (TP) = 26261	False Positive (FP) = 473683
Niet GTV	False Negative (FN) = 0	True Negative (TN) = 1138456

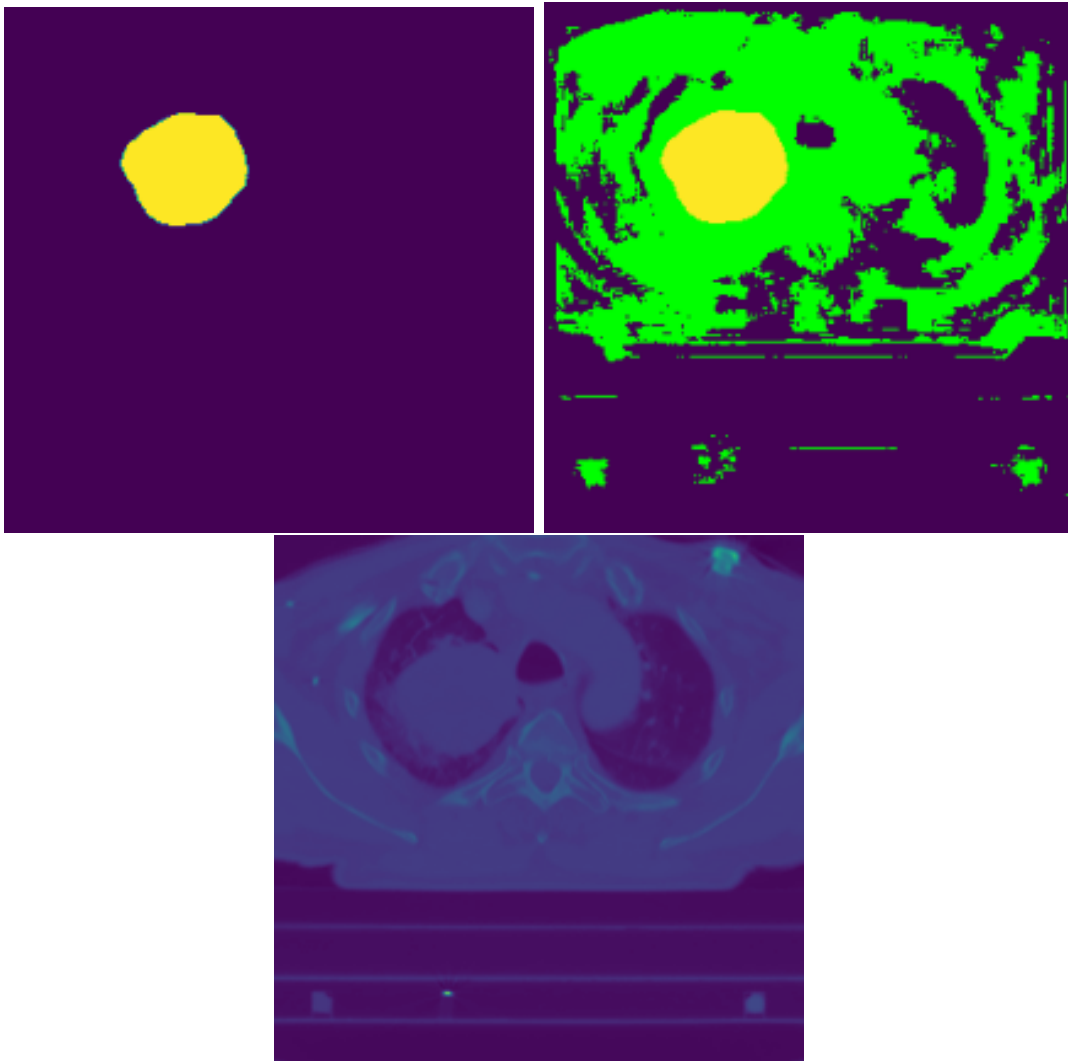
Indien we deze gegevens verwerken bekommen we volgende kwantitatieve indicatoren:

Tabel 27: True positive rate (TPR), true negative rate (TNR), false positive rate (FPR) en false negative rate (FNR) voor de test- en trainpatiënt na 100 epochs.

Patient	TPR = TP/P	TNR = TN/N	FPR = FP/N	FNR = FN/P
Test	0.996	0.639	0.361	0.004
Train	1.000	0.642	0.358	0.000

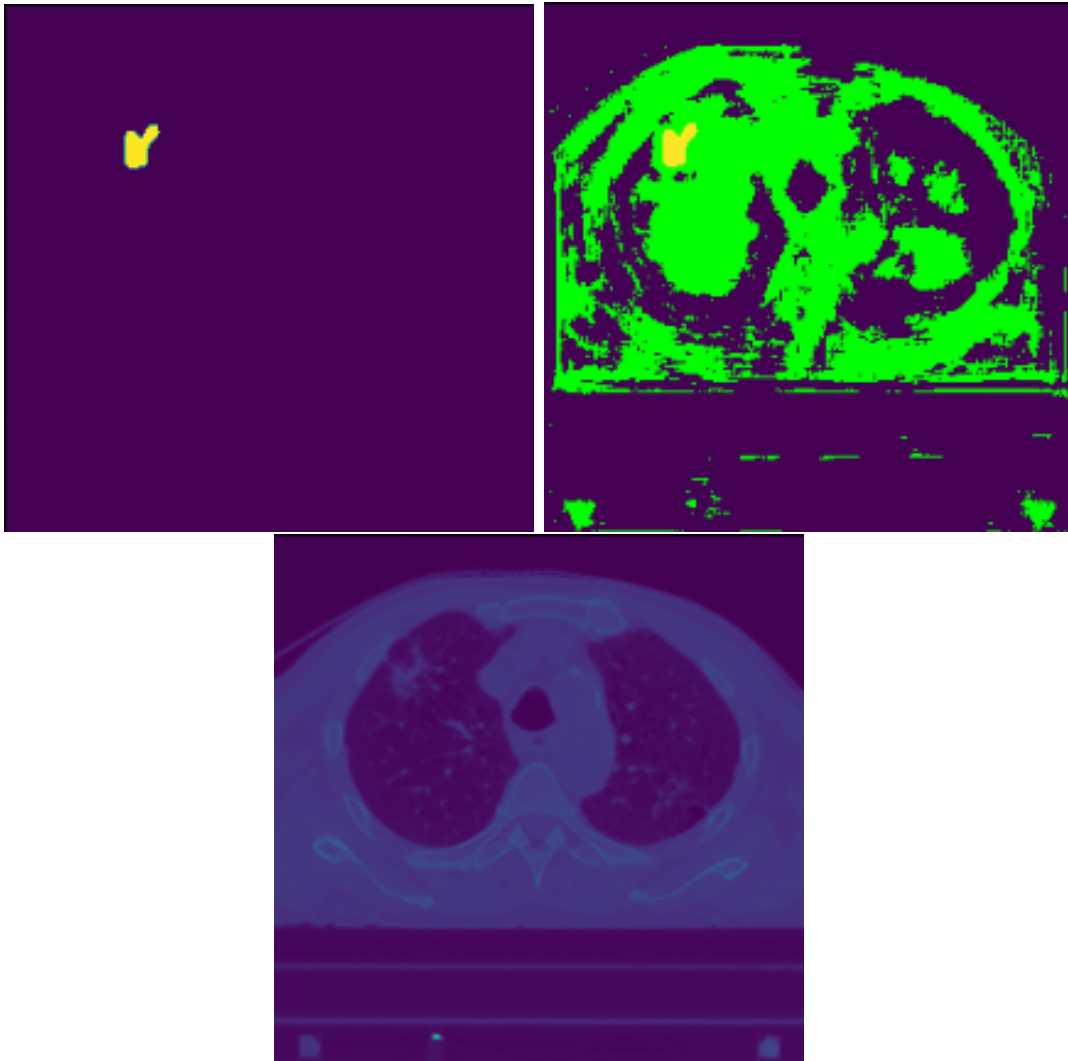
We merken zeer hoge TPR maar ook zeer hoge FPR en zeer lage TNR. Dit wijst naar een model dat veel meer GTV voorspelt dan er echt is.

Om ons vermoeden van een te grote voorspelling van GTV te staven kijken we weer naar enkele visuele voorspellingen. Paars is TN, geel is TP, groen is FP en rood is FN. De eerste set is afkomstig van de trainpatiënt en de tweede set van de testpatiënt.



Figuur 35: CT-beeld (onder) ground truth GTV-segmentatie (linksboven) en voorspelling experiment (rechtsboven) van de trainpatiënt.

Ons vermoeden is correct. Een aanzienlijk deel van de foto wordt onterecht voorspeld GTV te zijn. Dit is zeer slecht binnen de radiotherapie aangezien de locatie van de GTV sterk bepaalt waar en hoe sterk er bestraald wordt.



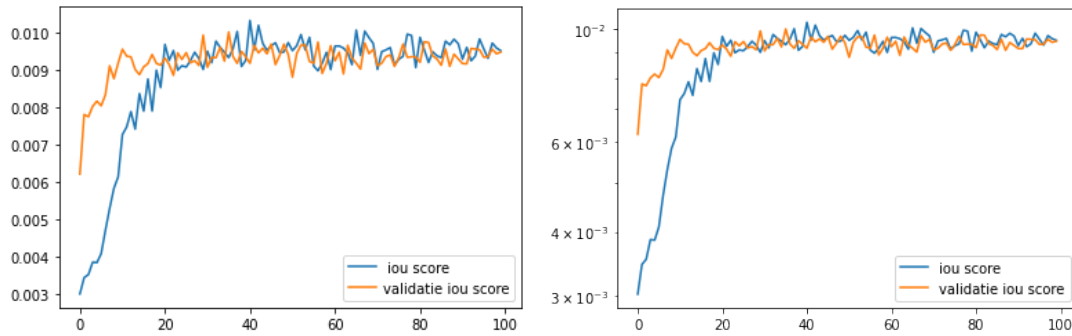
Figuur 36: CT-beeld (onder) ground truth GTV-segmentatie (linksboven) en voorspelling experiment (rechtsboven) van de testpatiënt.

Ook bij de testpatiënt, de patiënt waar het model van leert, zien we een ontoelaatbaar FPR.

Voor het GTV hadden we niet meteen een hoge verwachting. Het detecteren van tumoren is namelijk zeer anders van aard. De data is veel diverser en in dat opzicht is het ook meer een “outlier” detectieprobleem. Met deze methode zou er op zijn minst getraind moeten worden op een zeer grote en diverse dataset. Sowieso is onze manier om de gewichten te bepalen zoals besproken in sectie 2.2.2 niet goed gestaafd voor tumoren. Per lichaam verwacht men ongeveer hetzelfde percentage aan volume per orgaan. Dit kan men niet zeggen van tumoren.

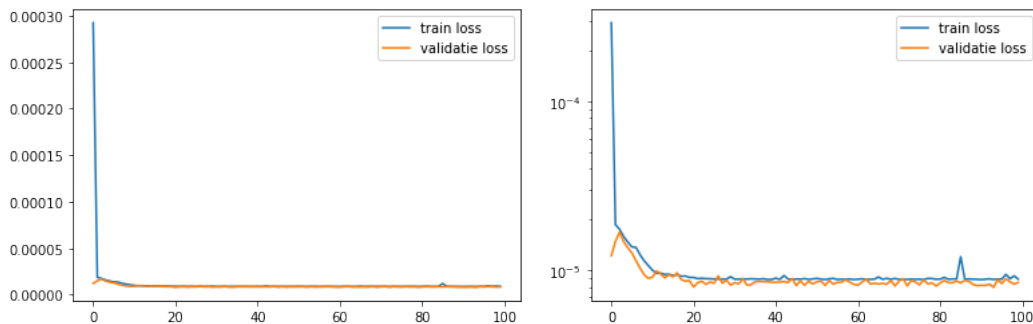


Zoals aangehaald voerden we voor het GTV een tweede test uit. Weer op drie schijnpatiënten voor 100 epochs maar deze keer met alle foto's, niet alleen de geannoteerde. Zo hopen we dat het model beter in staat is in te zien wanneer iets “niet tumor” is.



Figuur 37: Verloop van de IoU score op de trainset en de validatieset tijdens het trainen van het GTV over een periode van 100 epochs. Lineaire schaal (links) en logschaal (rechts).

We merken dat IoU score veel lager ligt dan bij het vorige experiment. Vermoedelijk zijn er nog steeds een hoop valse positieve die nu zwaarder doorwegen omdat er nu meer foto's waren waar er niks op geannoteerd staat. De evolutie van de IoU score heeft wel een meer straightforward karakter ten opzichte van het vorige experiment. Het model lijkt (over)traind te zijn. Dit is weer gekarakteriseerd door de globale stagnatie met de fluctuaties. Voor meer duidelijk kijken we verder naar de losswaarden:



Figuur 38: Verloop van de losswaarde op de trainset en de validatieset tijdens het trainen van het GTV over een periode van 100 epochs. Lineaire schaal (links) en logschaal (rechts).

De losswaarden lijken ook eerder een stagnatie te impliceren wat hint naar een (over)trainbaar model. Een (over)trainbaar model met slechte metrieken blijft echter een slecht model. We kijken verder naar de kwantitatieve metrieken.

De maximale train IoU score bedroeg 0.010 terwijl de test IoU score 0.001 bedroeg. De minimale train loss bedroeg 0.000009 terwijl de test loss 0.000012 bedroeg. Deze bevindingen wijzen sterk in het overtraind zijn van het model. Weer gaat de redenering op dat een overtrainbaar model met slechte trainpatiënt-resultaten nog steeds niet bruikbaar is.

We kijken verder naar de confusionmatrices om te zien of de TNR verbeterd is.

Tabel 28: Confusionmatrix van de testpatiënt na 100 epochs op het GTV.

Voorspelling/Ground Truth	GTV (P = 4768)	Niet GTV(N = 17296736)
GTV	True Positive (TP) = 4768	False Positive (FP) = 3297847
Niet GTV	False Negative (FN) = 0	True Negative (TN) = 13998889

Tabel 29: Confusionmatrix van de trainpatiënt na 100 epochs op het GTV.

Voorspelling/Ground Truth	GTV (P = 26261)	Niet GTV(N = 17406315)
GTV	True Positive (TP) = 26261	False Positive (FP) = 3035486
Niet GTV	False Negative (FN) = 0	True Negative (TN) = 14370829

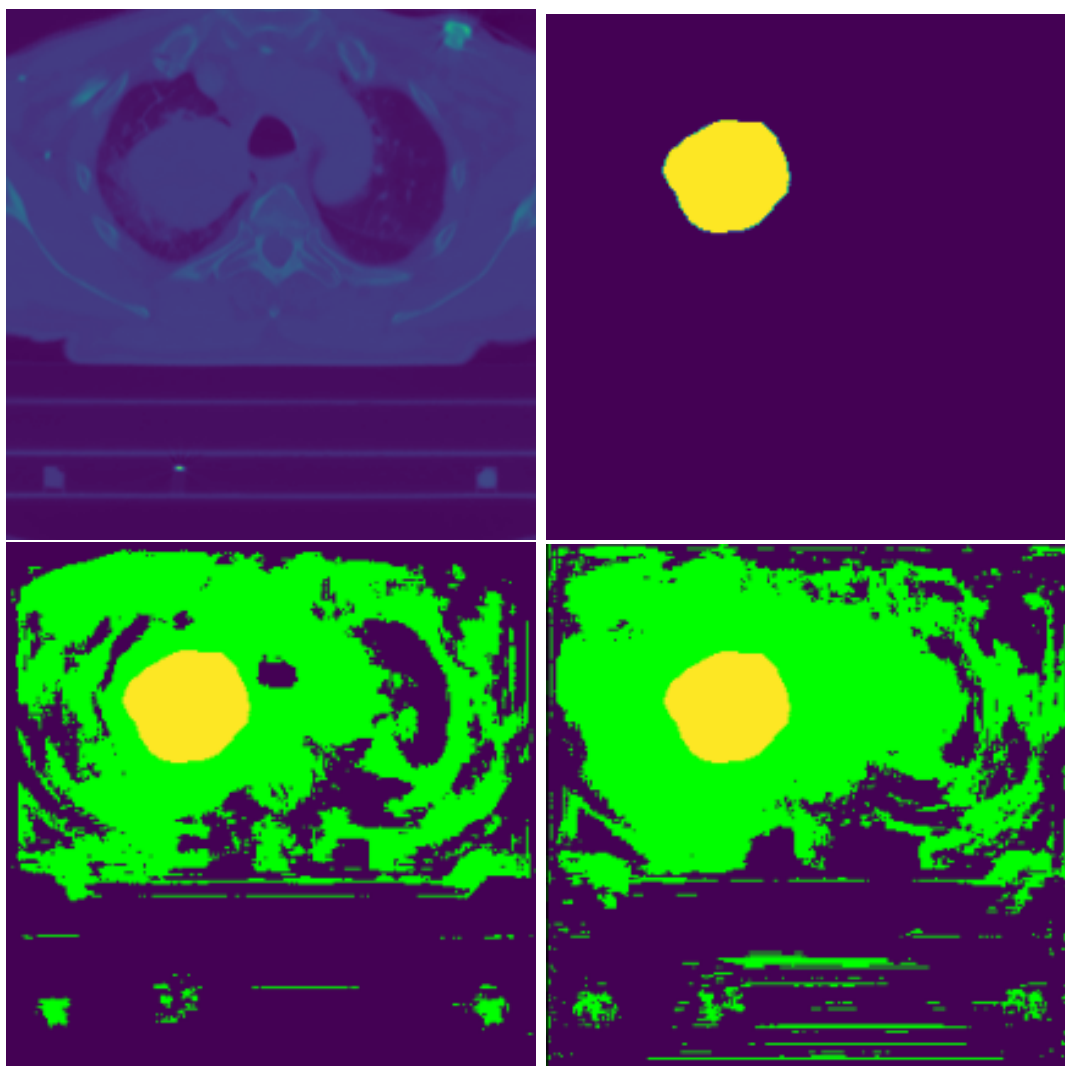
Indien we deze gegevens verwerken bekomen we volgende kwantitatieve indicatoren:

Tabel 30: True positive rate (TPR), true negative rate (TNR), false positive rate (FPR) en false negative rate (FNR) voor de test- en trainpatiënt na 100 epochs.

Patient	TPR = TP/P	TNR = TN/N	FPR = FP/N	FNR = FN/P
Test	1.000	0.809	0.191	0.000
Train	1.000	0.826	0.174	0.000

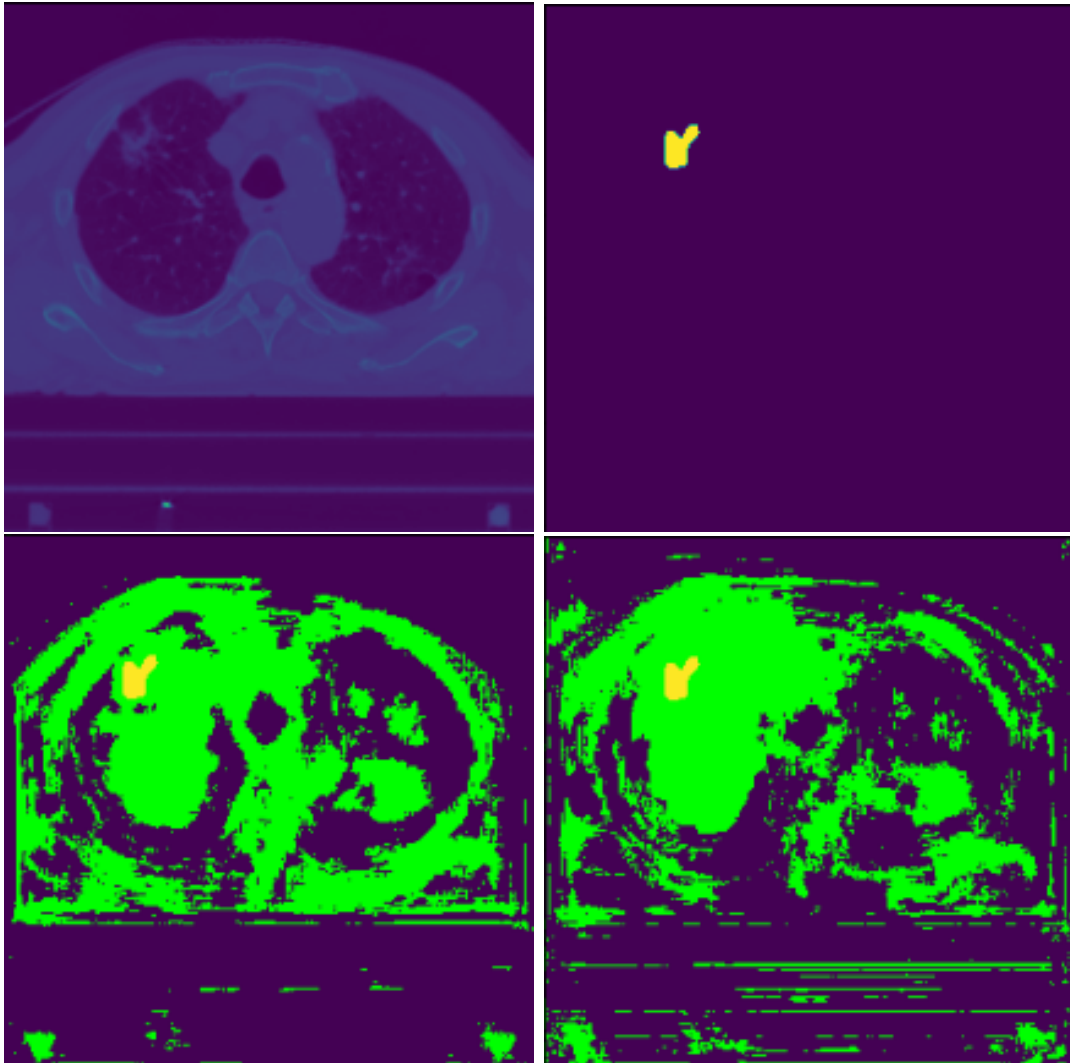
Zowel bij de testpatiënt als de trainpatiënt hebben we nu een perfecte TPR. De TNR ligt echter nog steeds beduidend laag. Er mag niet veel toegeving zijn op de TNR en FPR, want men wilt zaken die geen GTV zijn niet classificeren als GTV.

Om de verbetering in de TNR en FPR te verklaren kijken we naar de foto's. Paars is TN, geel is TP, groen is FP en rood is FN. De eerste set is afkomstig van de trainpatiënt en de tweede set van de testpatiënt.



Figuur 39: CT-beeld (linksboven) ground truth GTV-segmentatie (rechtsboven) voorspelling experiment 1 (linksonder) en voorspelling experiment 2 (rechtsonder) van de trainpatiënt.

Uit deze beelden zien we geen sterke verbetering van dit experiment ten opzichte van het vorige experiment. Zelfs niet voor de trainpatiënt.



Figuur 40: CT-beeld (linksboven) ground truth GTV-segmentatie (rechtsboven) voorspelling experiment 1 (linksonder) en voorspelling experiment 2 (rechtsonder) van de testpatiënt.

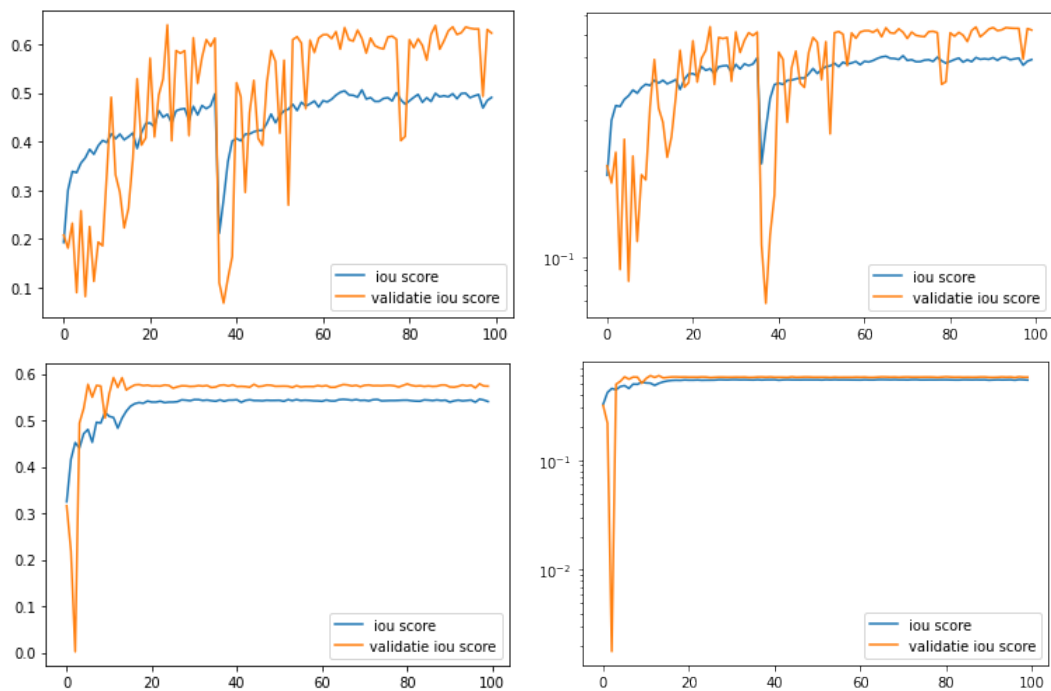
Ook voor de testpatiënt zien we geen overduidelijke visuele verbetering van de voorspellingen.

We kunnen nog steeds geen conclusie trekken of dit type model trainbaar is voor tumoren of niet. We zouden veel meer beelden zonder tumoren maar ook veel meer beelden met tumoren nodig hebben. Zo kan het model enerzijds beter leren hoe de menselijke anatomie er uit hoort te zien en anderzijds beter de grote variatie aan tumoren leren herkennen.

### 3.3 Gewogen categoriale crossentropie loss

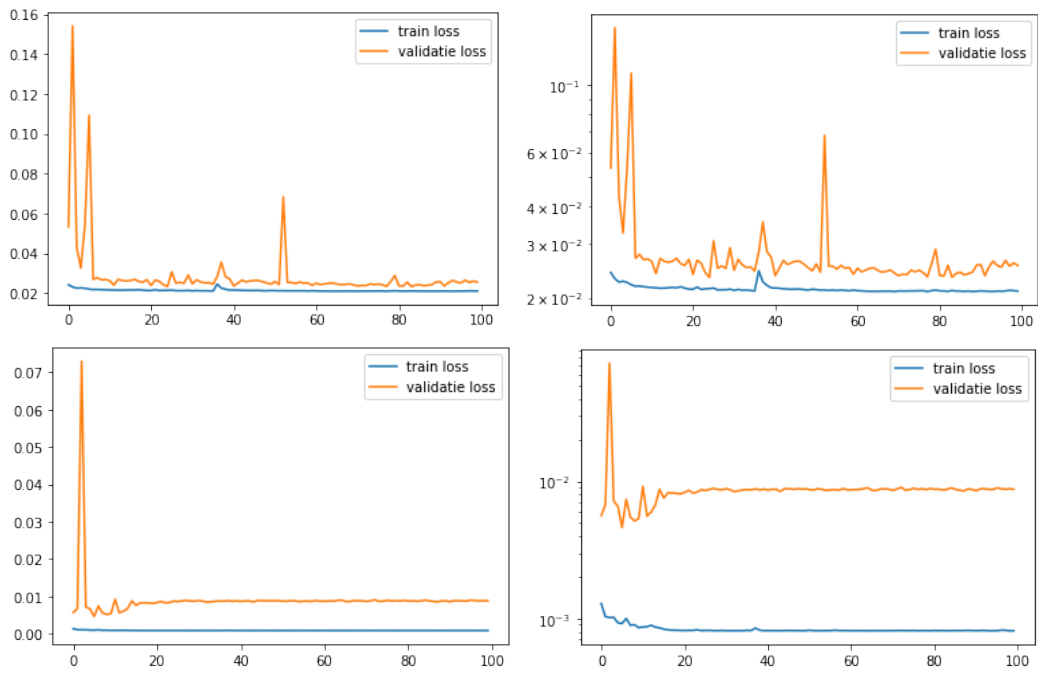
Hier werden alle organen tegelijk getraind met de categoriale crossentropie loss zoals beschreven in sectie 2.2.3 over een periode van 100 epochs. Eerst op de configuratie met drie patiënten en daarna op de configuratie met 29 patiënten. Als metriek gebruiken we de IoU score zoals beschreven in sectie 2.4.2. We zijn met deze test vooral geïnteresseerd in de verschillen tussen het trainen op de kleine groep en het trainen op de grote groep. Achteraf zullen we oordelen of deze lossfunctie al dan niet in staat is om gebruikt te worden voor het trainen van het voor ons wenselijk model.

We zullen eerst de traingrafieken, IoU score en gewogen Diceloss tussen configuratie 1 en configuratie 2 vergelijken waarbij we ons telkens beperken tot dezelfde train- en testpatiënt. Daarna zullen we voor elk orgaan apart de IoU score, Diceloss en voorspellingen vergelijken waarbij we weer dezelfde train- en testpatiënt gebruiken. Een beknopte bespreking bevindt zich in sectie 4.3.



Figuur 41: Verloop van de IoU score op de trainset en de validatieset tijdens het trainen van alle organen over een periode van 100 epochs. Lineaire schaal (links) en logschaal (rechts). Configuratie 1 (boven) en configuratie 2 (onder).

We zien bij configuratie 1 dat de de validatie score al snel boven de test score komt te liggen. Dit is merkwaardig en doet ons vermoeden dat er een grote statistische fout op de waarden zit. We zouden echter verwachten dat de fout vooral groot is in het begin en afneemt naar het einde toe. Dit zien we niet bij configuratie 2. Daar zien we al vlug een stabilisatie van de scores ten gevolge van de grotere groepen data die het model krijgt per epoch. Ook daar blijft de validatie score consistent boven de test score liggen. De waarde van de IoU is wat we verwachten aangezien het het uitgemiddelde resultaat zou moeten zijn van de binaire IoU scores.



Figuur 42: Verloop van de losswaarde op de trainset en de validatieset tijdens het trainen van alle organen over een periode van 100 epochs. Lineaire schaal (links) en logschaal (rechts). Configuratie 1 (boven) en configuratie 2 (onder).

De grafieken van de losswaarden zijn zeer gelijkend. Er zijn hoge pieken in het begin die min of meer vlug stabiliseren. Bij configuratie 1 zijn de fluctuaties iets meer volatiel. Opvallend is wel dat de kloof tussen test en validatie groter is bij configuratie 2 dan bij configuratie 1. Dit is vreemd aangezien we verwachten dat configuratie 2 (met meer data) beter zou zijn in het veralgemeniseren.

Tabel 31: Dicescores van de testgroep van configuratie 1 en configuratie 2.

Metriek	conf1	conf2
test IoU	0.587	0.557
test loss	0.0208	0.0009
max train IoU	0.506	0.546
min train loss	0.0209	0.0008

We zien een verbetering in de maximale train IoU van configuratie 1 naar configuratie 2. Dit is wenselijke aangezien dit betekent dat het model schaalbaar is naar grotere datasets. We zien echter wel een daling in de test IoU. Dit doet vermoeden dat configuratie 2 meer overtraint is dan configuratie 1. Dit is mogelijk indien de grotere dataset alsnog redelijk homogeen is, want per epoch ziet het model zo veel meer data. We zien ook dat de test IoU consistent hoger is dan de maximale train IoU. Dit is merkwaardig aangezien we het omgekeerde verwachten wat weer zou suggereren dat er redelijke fouten zitten op deze metriek.

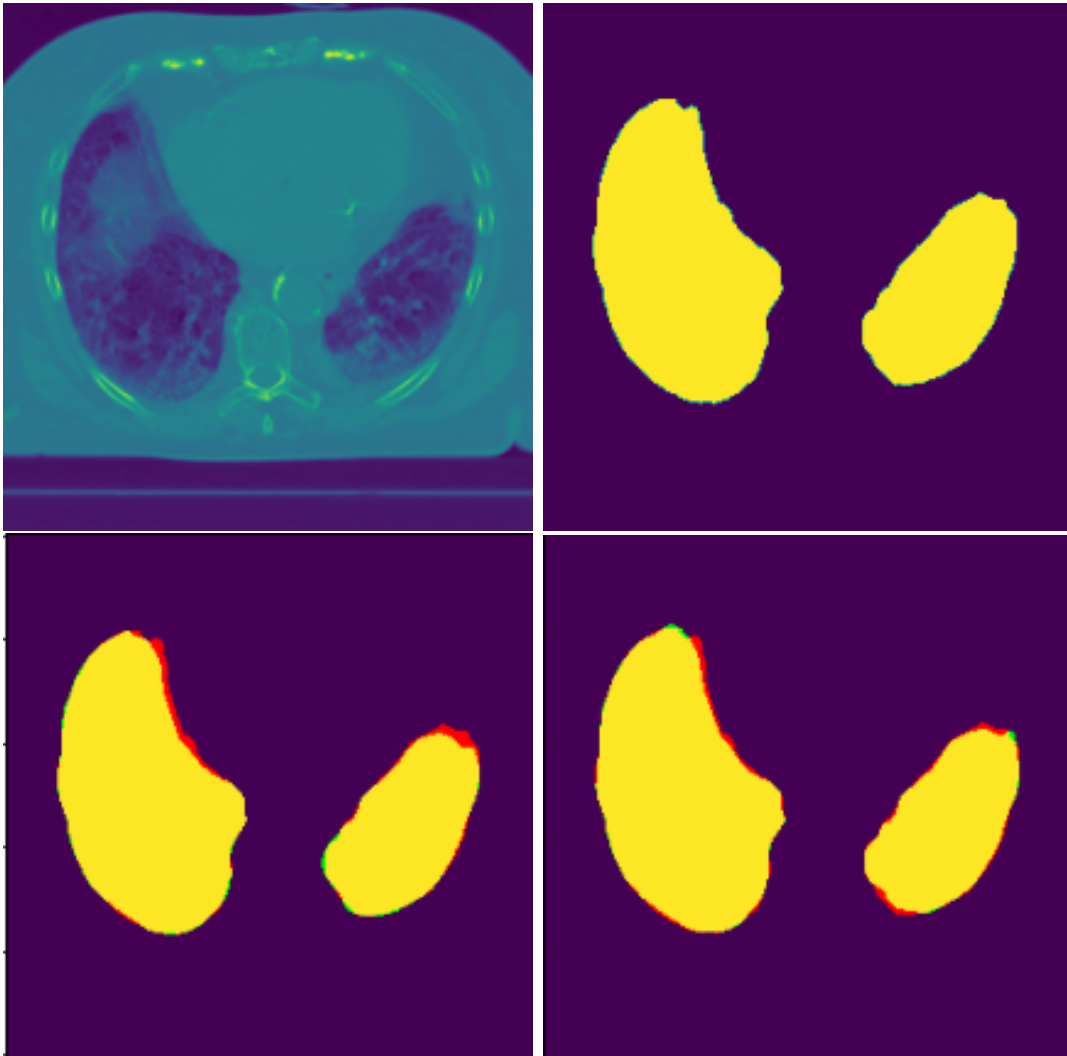
Tabel 32: IoU scores per orgaan van de testgroep van configuratie 1 en configuratie 2.

Orgaan	conf1	conf2
Hart	0.674	0.776
Slokdarm	0.300	0.272
GTV	0.074	0.184
Longen	0.926	0.944
luchtpijp	0.603	0.587
Ruggenmerg	0.497	0.453

We zien dat de helft van de organen is verbeterd en de helft is verslechterd bij het toevoegen van meer data. We kijken nu per orgaan naar de voorspellingen om te zien hoe exact het model verbeterd of verslechterd is van configuratie 1 naar 2.

### 3.3.1 Longen

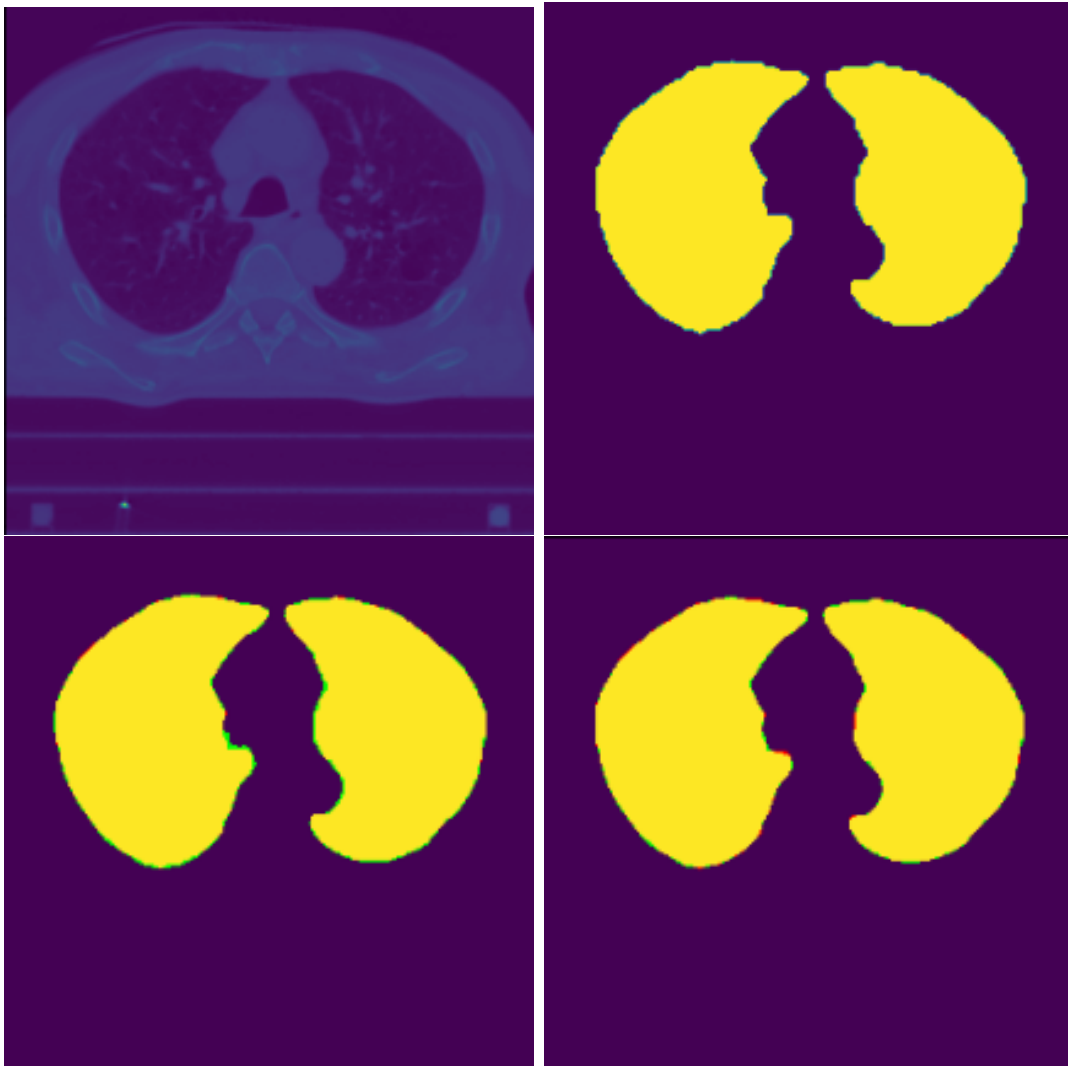
Paars is TN, geel is TP, groen is FP en rood is FN. De eerste set is afkomstig van de trainpatiënt en de tweede set van de testpatiënt.



Figuur 43: CT-beeld (linksboven) ground truth long-segmentatie (rechtsboven) voorspelling configuratie 1 (linksonder) en voorspelling cofiguratie 2 (rechtsonder) van de trainpatiënt.

We zien zeer sterke overeenkomst tussen het ware masker en de voorspellingen door beide configuraties. Ook komen beide voorspellingen van de twee configuraties zeer goed overeen. Dit is nodig aangezien het falen van het trainen van de longen zou betekenen dat het model zeker niet trainbaar is gezien de lage moeilijkheidsgraad van de longen.





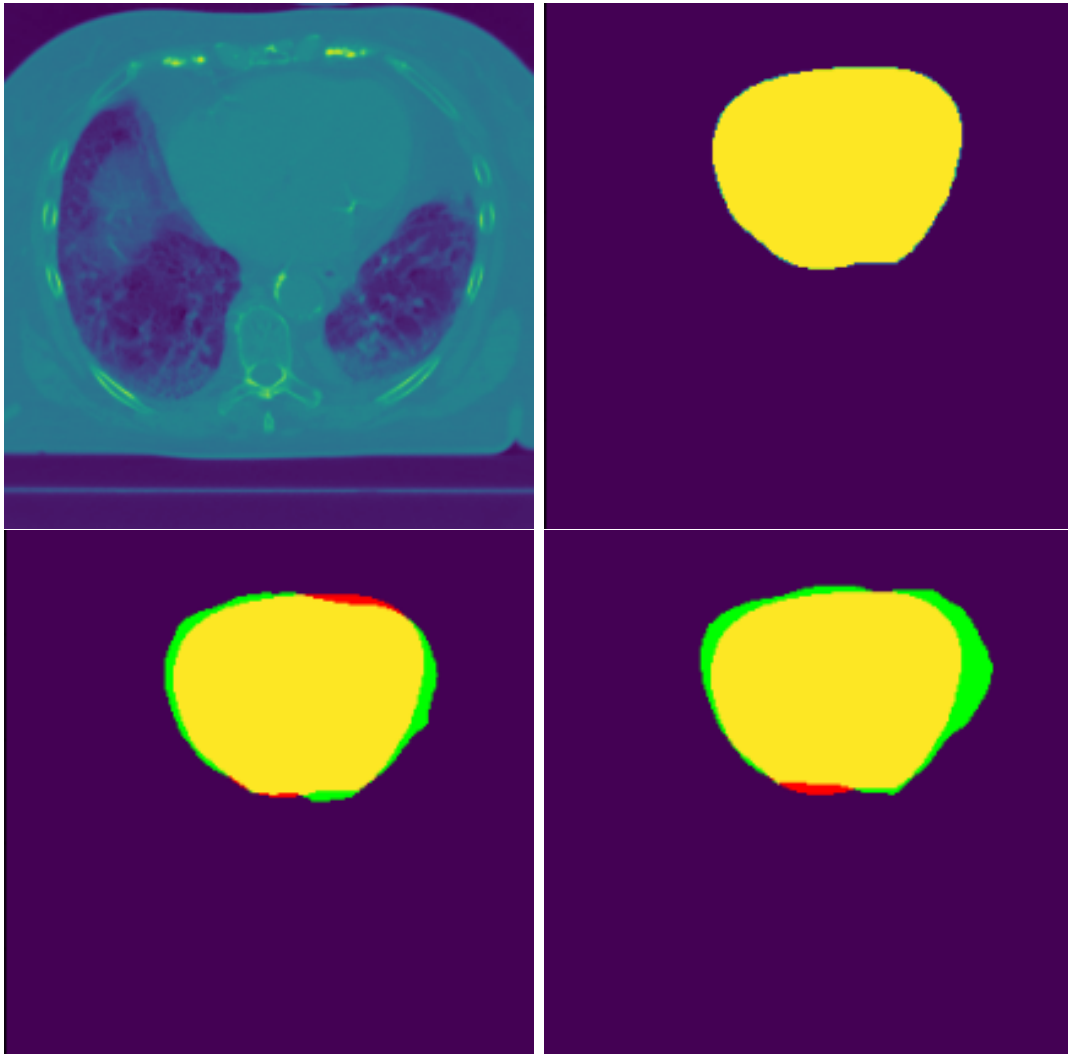
Figuur 44: CT-beeld (linksboven) ground truth long-segmentatie (rechtsboven) voorspelling configuratie 1 (linksonder) en voorspelling configuratie 2 (rechtsonder) van de testpatiënt.

Opnieuw zien we een bijna perfecte gelijkheid tussen het ware masker en de voorspellingen van beide configuraties.

Algemeen kunnen we concluderen dat de categoriale aanpak in staat is voldoende correcte voorspellingen te maken over de longen.

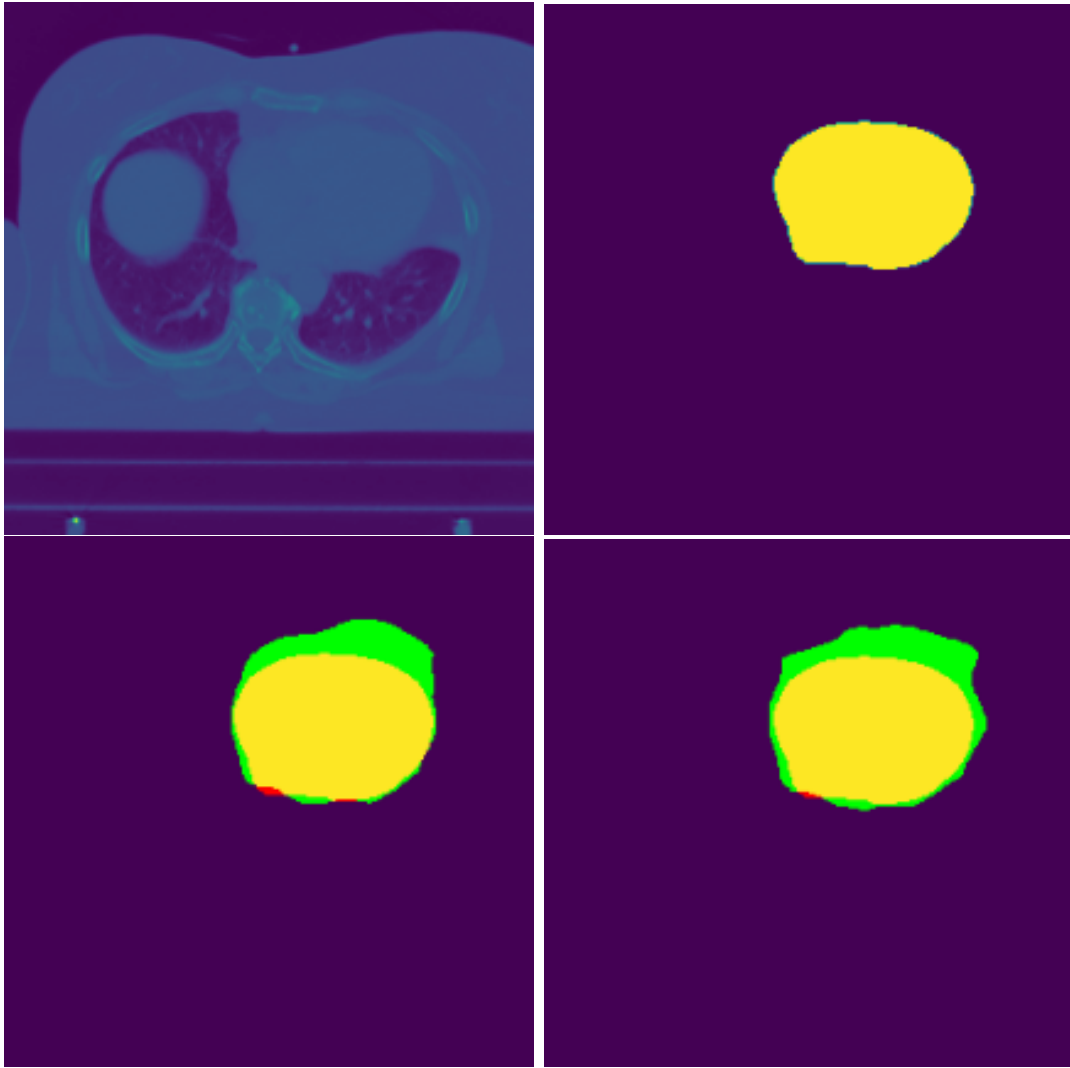
### 3.3.2 Hart

Paars is TN, geel is TP, groen is FP en rood is FN. De eerste set is afkomstig van de trainpatiënt en de tweede set van de testpatiënt.



Figuur 45: CT-beeld (linksboven) ground truth hart-segmentatie (rechtsboven) voorspelling configuratie 1 (linksonder) en voorspelling configuratie 2 (rechtsonder) van de trainpatiënt.

We zien vooral valse positieven en, relatief gezien, kleine valse negatieven. Opmerkelijk is dat de situatie verslechterd lijkt te zijn bij het toenemen van de trainingsdata. Het aantal fouten is algemeen toegenomen. De valse negatieven zijn echter afgenomen. Dit is wenselijker bij het intekenen van een orgaan. De valse positieven bij configuratie 2 zijn echter relatief groot ten opzichte van het orgaan. Dit is iets wat nog verder verbeterd moet worden.



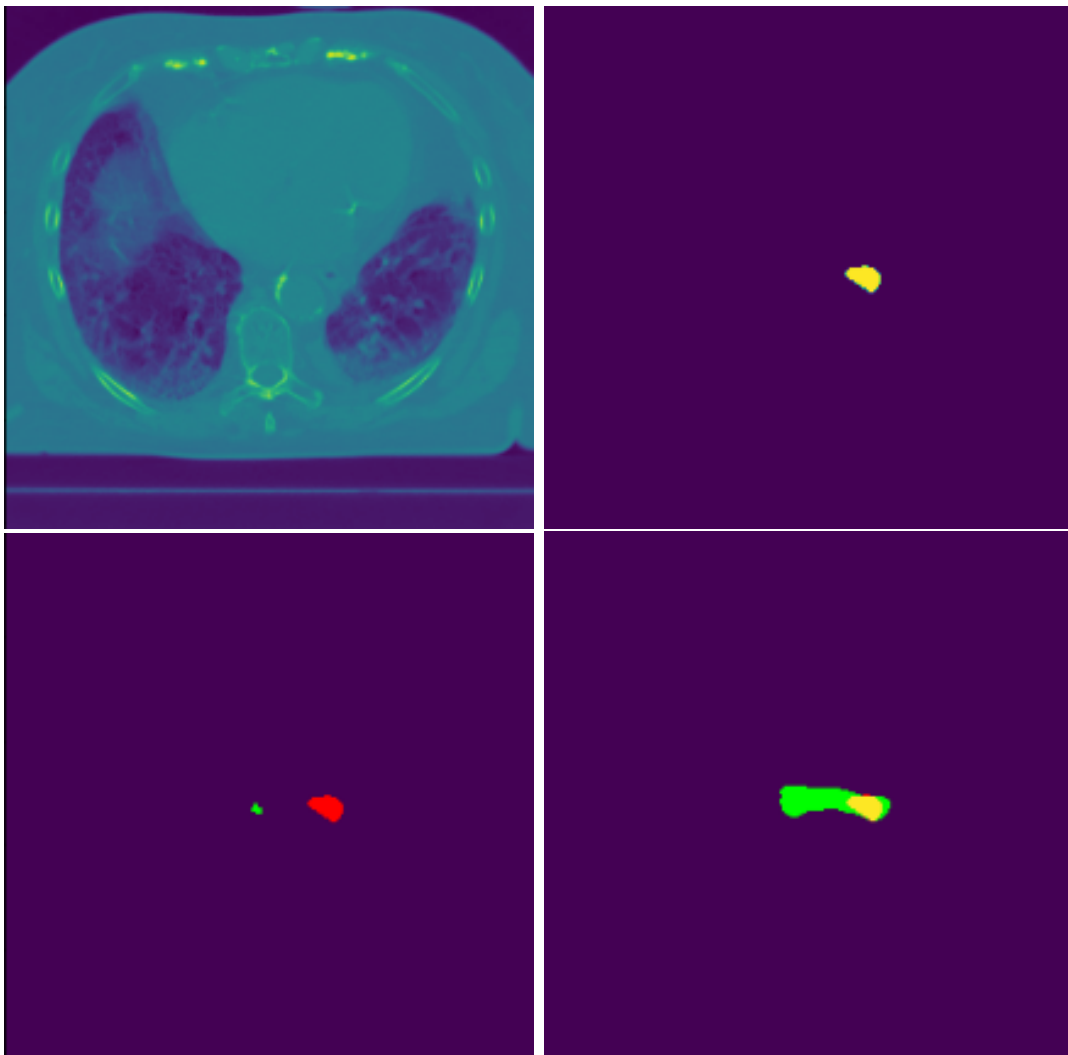
Figuur 46: CT-beeld (linksboven) ground truth hart-segmentatie (rechtsboven) voorspelling configuratie 1 (linksonder) en voorspelling configuratie 2 (rechtsonder) van de testpatiënt.

De valse positieven bij de testpatiënt zijn wel uitzonderlijk groot. Alhoewel dit minder erg is dan valse negatieven bij een orgaan, is deze orde van fout wel extreem.

We concluderen dat het categoriaal trainen op het hart nog niet compleet wenselijk is, maar dat verdere verfijningen van het model, zoals aan de gewichten, nodig zijn om definitief de effectiviteit van het model te bepalen.

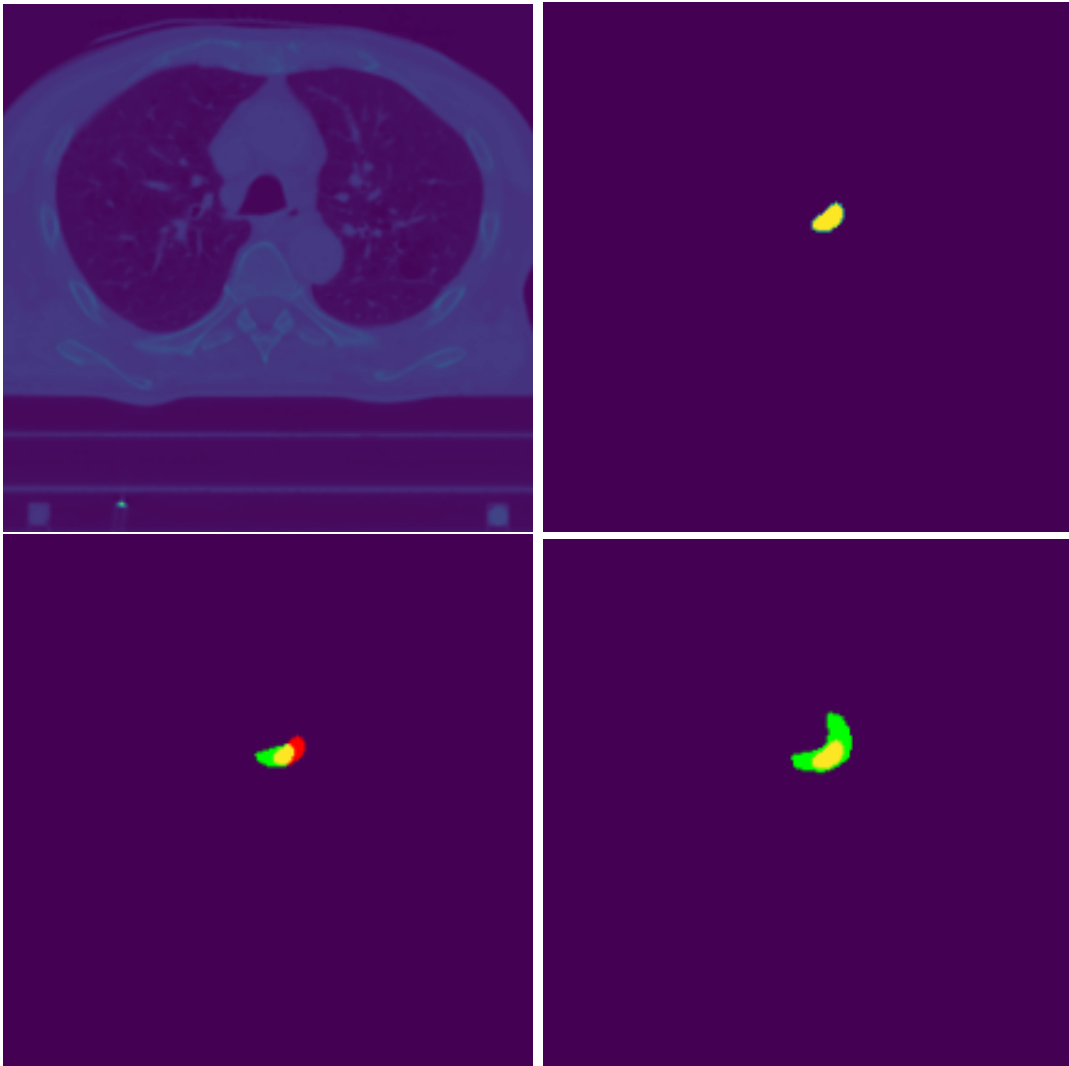
### 3.3.3 Slokdarm

Paars is TN, geel is TP, groen is FP en rood is FN. De eerste set is afkomstig van de trainpatiënt en de tweede set van de testpatiënt.



Figuur 47: CT-beeld (linksboven) ground truth slokdarm-segmentatie (rechtsboven) voorspelling configuratie 1 (linksonder) en voorspelling configuratie 2 (rechtsonder) van de trainpatiënt.

Bij configuratie 1 was het model volledig niet in staat het orgaan terug te vinden. Dit is het slechtst mogelijke scenario. De stralingsplanningsoftware zou op deze manier geen rekening houden met dit orgaan. Bij configuratie 2 zien we dat het model min of meer het orgaan terugvindt, maar dit ook zeer sterk overschat. Dergelijke fouten zijn echter minder erg dan helemaal niks terug te vinden, maar nog steeds ver van aanvaardbaar.



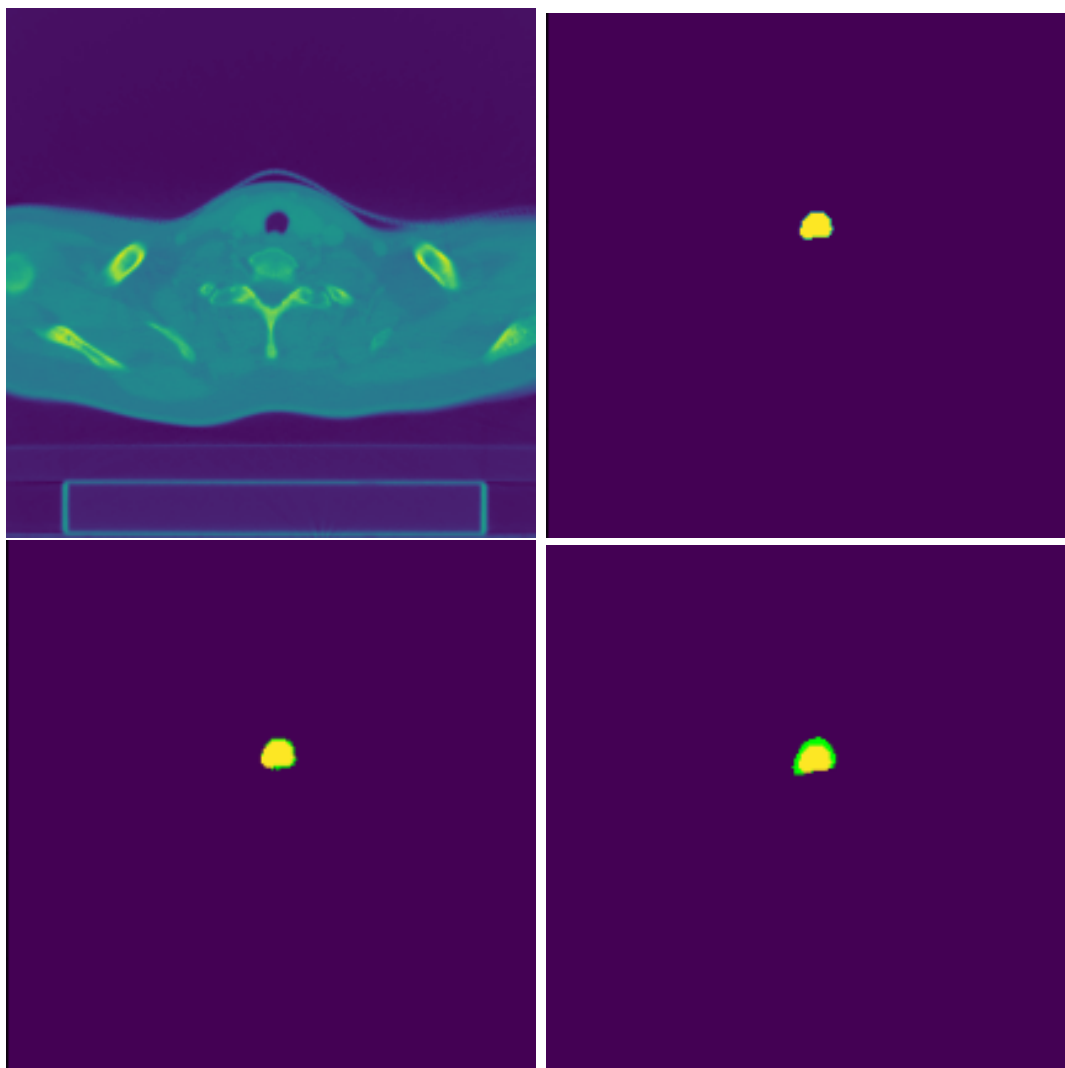
Figuur 48: CT-beeld (linksboven) ground truth slokdarm-segmentatie (rechtsboven) voorspelling configuratie 1 (linksonder) en voorspelling configuratie 2 (rechtsonder) van de testpatiënt.

Aangezien de situatie niet wenselijk was bij de trainpatiënt, is ze dit ook niet bij de testpatiënt. Opnieuw heeft configuratie 2 de drang om sterke overschattingen te maken.

We concluderen dat het categoriaal trainen op de slokdarm nog ver van wenselijk is, maar dat verdere verfijningen van het model, zoals aan de gewichten, nodig zijn om definitief de effectiviteit van het model te bepalen.

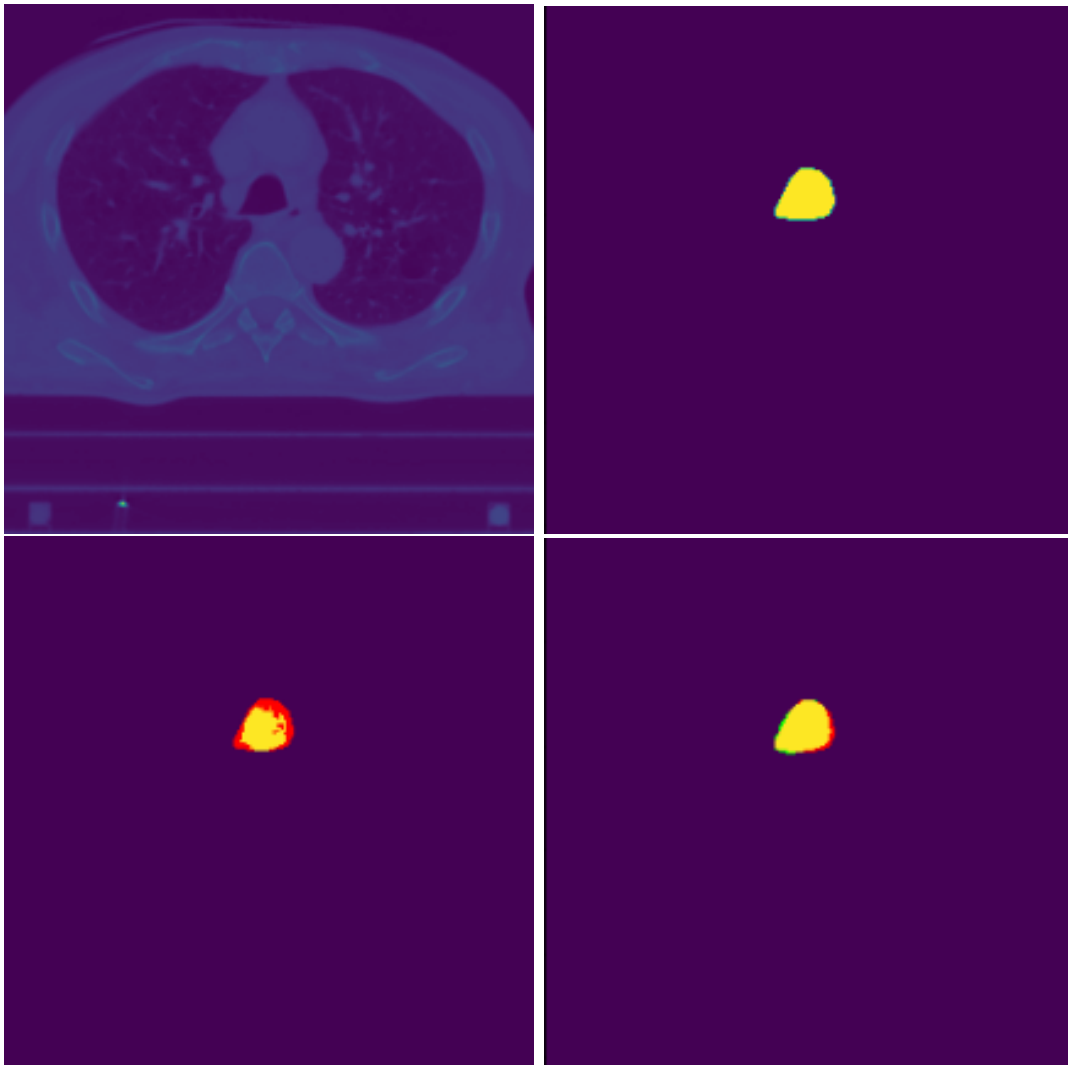
### 3.3.4 Luchtpijp

Paars is TN, geel is TP, groen is FP en rood is FN. De eerste set is afkomstig van de trainpatiënt en de tweede set van de testpatiënt.



Figuur 49: CT-beeld (linksboven) ground truth luchtpijp-segmentatie (rechtsboven) voorspelling configuratie 1 (linksonder) en voorspelling configuratie 2 (rechtsonder) van de trainpatiënt.

We zien dat het model redelijk goed in staat is de luchtpijp terug te vinden. Wat opvalt is dat de situatie bij configuratie 2 minder wenselijk is dan bij configuratie 1. We weten echter dat de IoU score groter is bij configuratie 2 dan bij 1. Vermoedelijk is deze situatie dus een uitschieter. Daarnaast zijn de fouten valse positieven, die wenselijker zijn dan valse negatieven. Ook zijn het relatief kleine fouten.



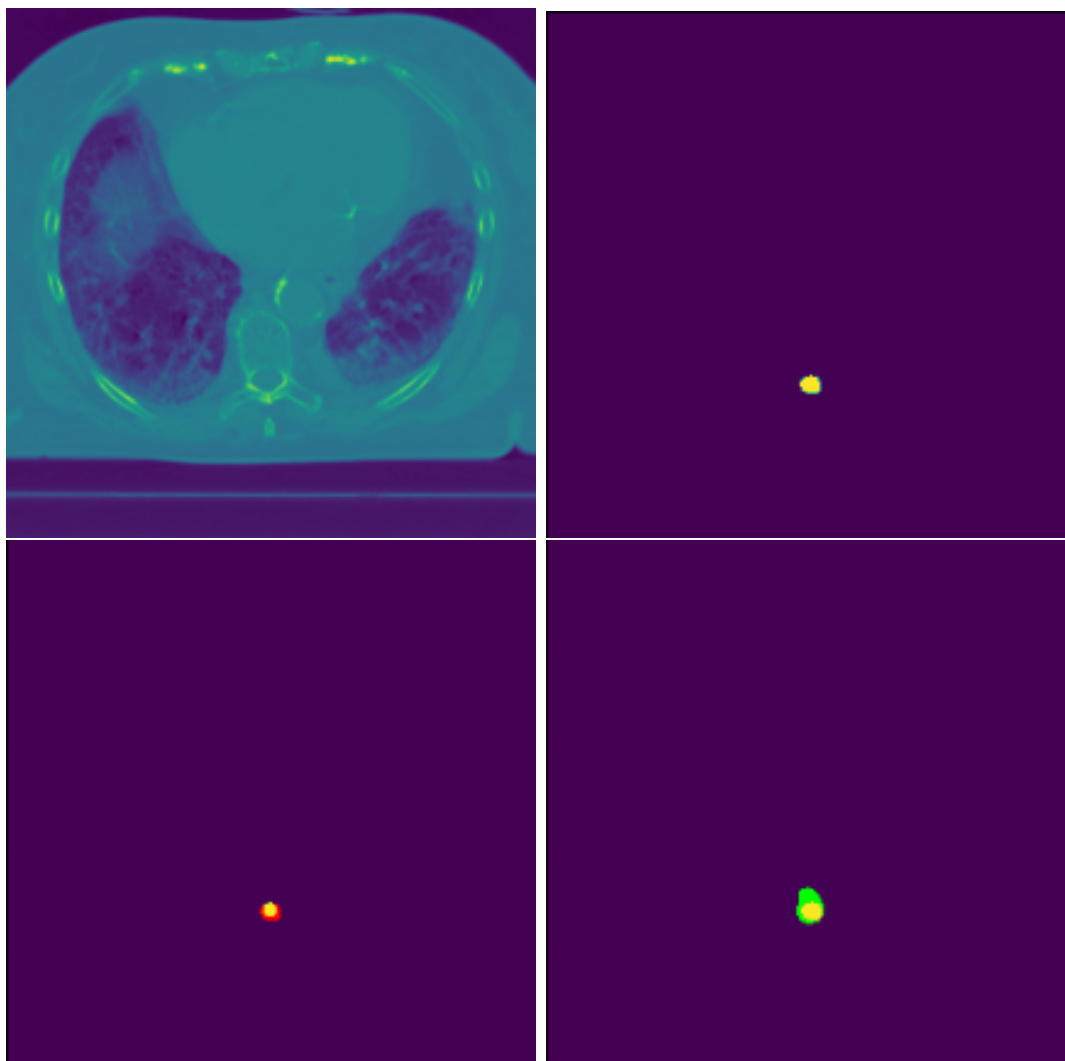
Figuur 50: CT-beeld (linksboven) ground truth luchtpijp-segmentatie (rechtsboven) voorspelling configuratie 1 (linksonder) en voorspelling configuratie 2 (rechtsonder) van de testpatiënt.

De testpatiënt doet het voor configuratie 1 slechter dan de trainpatiënt maar voor configuratie 2 beter dan de trainpatiënt. Dit staft verder ons vermoeden dat de vorige reeks foto's een uitschieter was. De voorspelling bij configuratie 2 is nagenoeg perfect met slechts kleine fouten aan de rand.

We concluderen dat het categoriaal trainen op de luchtpijp vrij wenselijk is, maar eventueel nog verfijnd kan worden.

### 3.3.5 Ruggenmerg

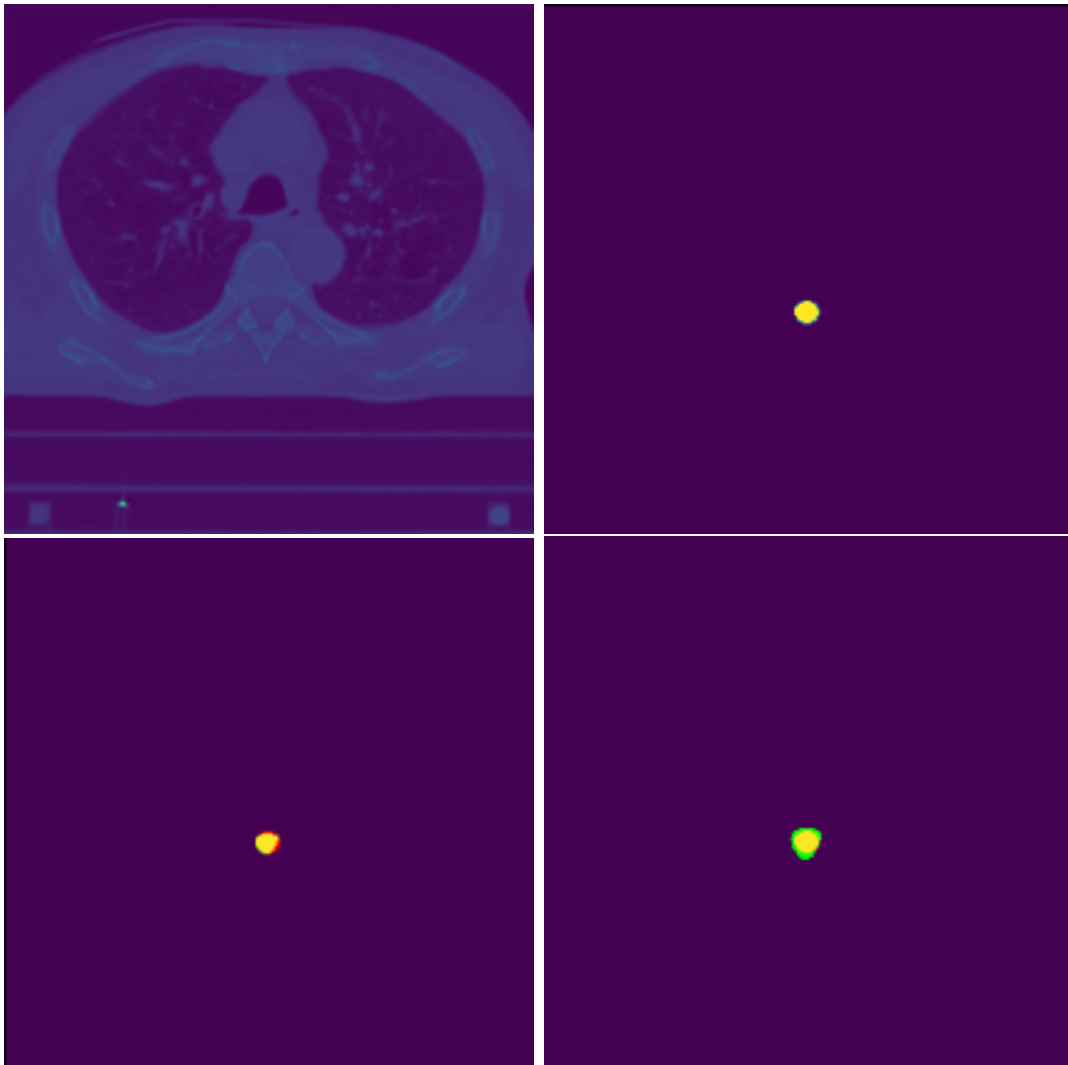
Paars is TN, geel is TP, groen is FP en rood is FN. De eerste set is afkomstig van de trainpatiënt en de tweede set van de testpatiënt.



Figuur 51: CT-beeld (linksboven) ground truth ruggenmerg-segmentatie (rechtsboven) voorspelling configuratie 1 (linksonder) en voorspelling configuratie 2 (rechtsonder) van de trainpatiënt.

We zien dat het model in beide situaties redelijk in staat is het orgaan terug te vinden. Situatie 2 heeft ten opzichte van situatie 1 meer valse positieven en minder valse negatieven. Dit is wenselijk aangezien het niet voldoende aflijnen van een orgaan erger is dan de grootte van het orgaan licht te overschatten. De fouten aan de rand zijn echter nog relatief groot ten opzichte van het echte orgaan.





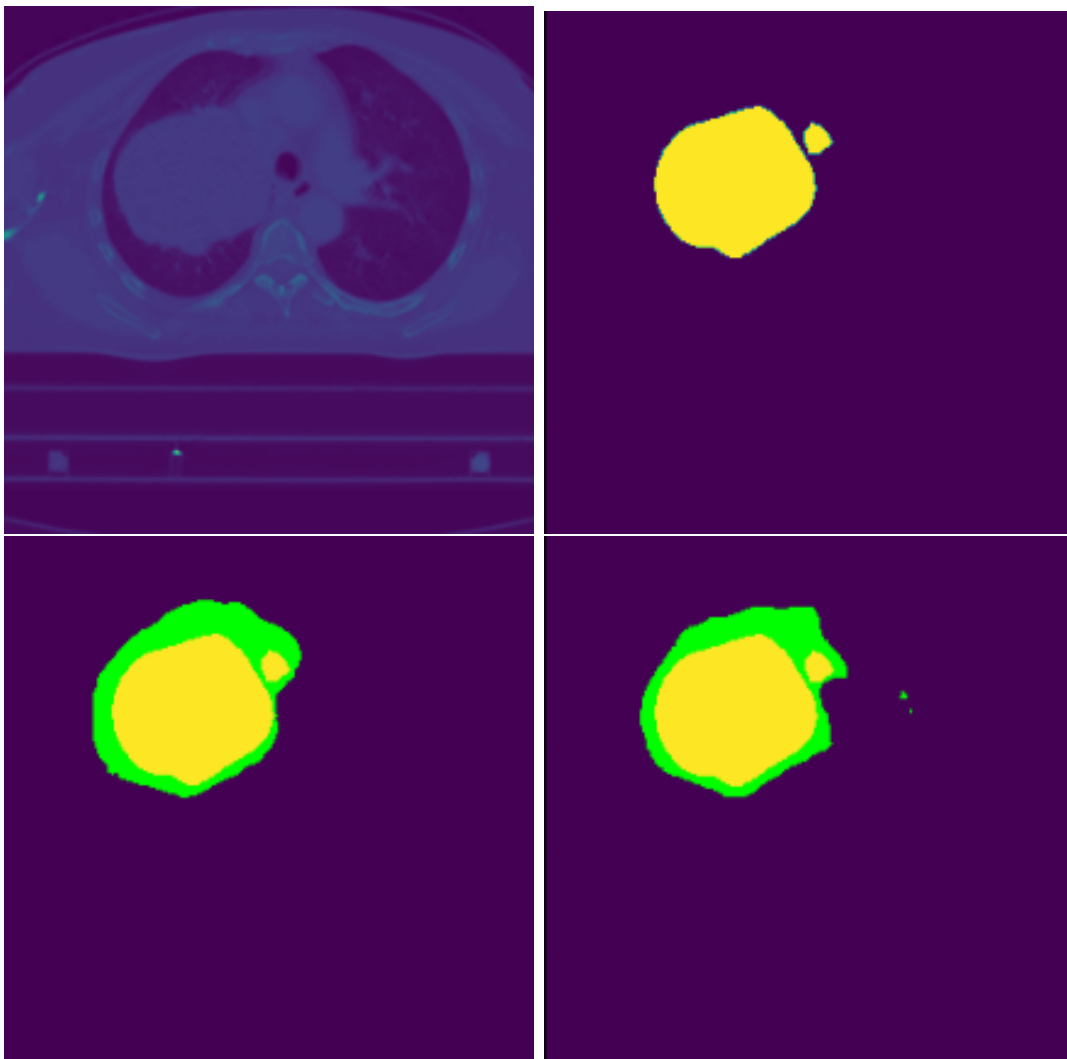
Figuur 52: CT-beeld (linksboven) ground truth ruggenmerg-segmentatie (rechtsboven) voorspelling configuratie 1 (linksonder) en voorspelling configuratie 2 (rechtsonder) van de testpatiënt.

Opnieuw lijkt de testpatiënt het iets beter te doen dan de trainpatiënt. De fouten zijn iets wenselijker aangezien de fouten relatief gezien niet zo extreem zijn ten opzichte van de ware grootte van het orgaan.

We concluderen dat het categoriaal trainen op het ruggenmerg vrij wenselijk is, maar eventueel nog verfijnd kan worden.

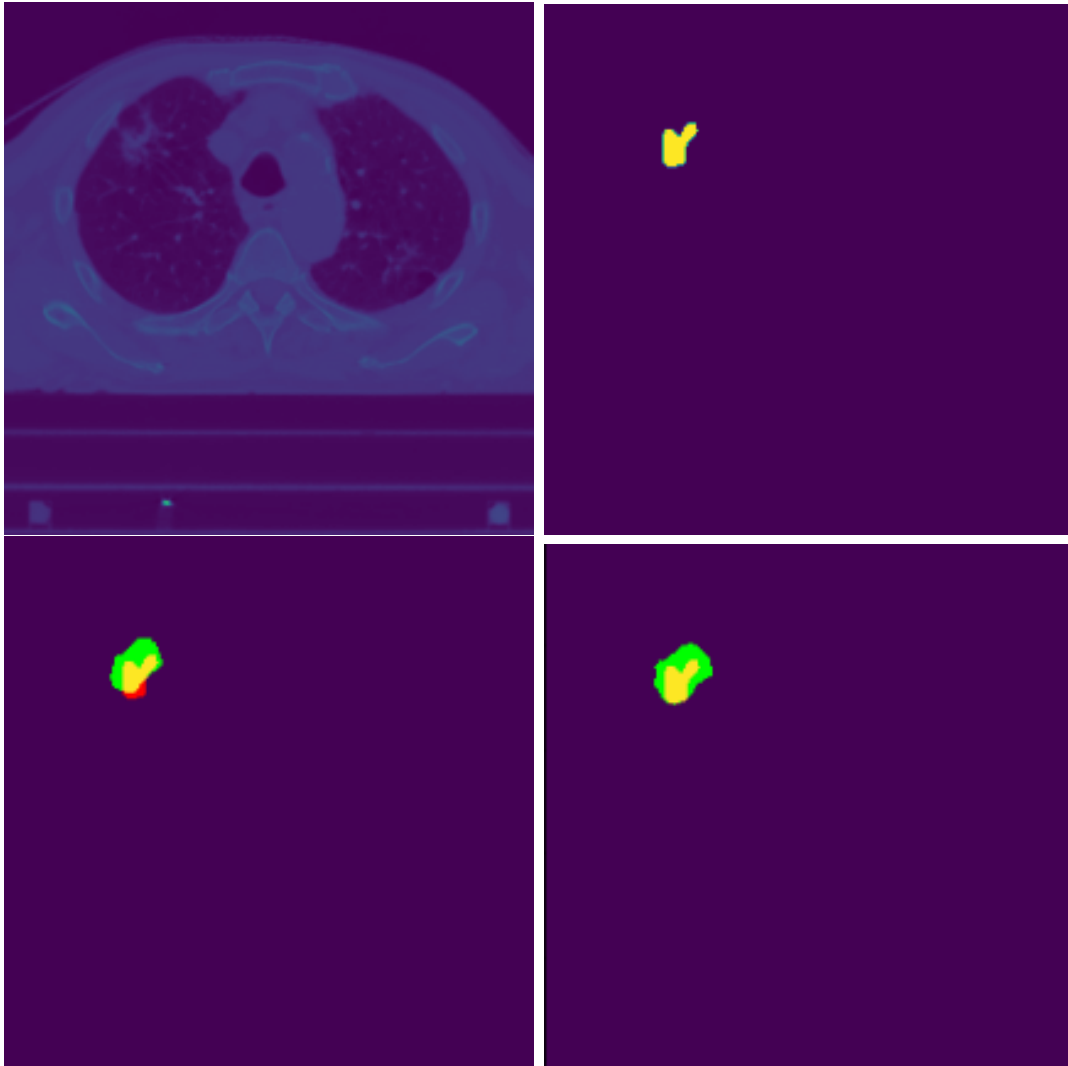
### 3.3.6 GTV

Paars is TN, geel is TP, groen is FP en rood is FN. De eerste set is afkomstig van de trainpatiënt en de tweede set van de testpatiënt.



Figuur 53: CT-beeld (linksboven) ground truth GTV-segmentatie (rechtsboven) voorspelling configuratie 1 (linksonder) en voorspelling configuratie 2 (rechtsonder) van de trainpatiënt.

Opmerkelijk is dat het model veel beter presteert ten opzichte van zijn binaire equivalent. Echter is er nog steeds een zeer groot gebied van valse positieven. Dit is zeker niet wenselijk aangezien het intekenen van de tumor bepalend is voor het doelgebied van de stralingsdosis.



Figuur 54: CT-beeld (linksboven) ground truth GTV-segmentatie (rechtsboven) voorspelling configuratie 1 (linksonder) en voorspelling configuratie 2 (rechtsonder) van de testpatiënt.

We zien hetzelfde als bij de trainpatiënt. Er wordt nog een relatief groot gebied onterecht aangeduid als tumor. Als gevolg zal de gekregen dosis onnodig hoog liggen. We maken wel weer de bemerking dat de situatie veel wenselijker is dan bij het binaire geval.

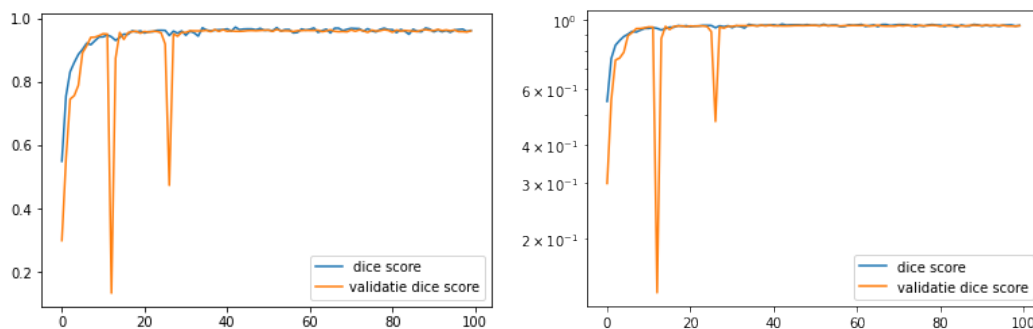
We concluderen dat het categoriaal trainen op het GTV potentie heeft aangezien het model redelijk in staat was het GTV terug te vinden wat opmerkelijk is gezien de aard van het GTV. Echter is de situatie nog ver van wenselijk aangezien er een relatief groot aantal valse positieven zijn. Verdere testen zijn nodig om te kijken of dit nog te verfijnen valt.

## 3.4 Dicescore per orgaan

Voor elk orgaan worden verschillende kwalitatieve en kwantitatieve tests gedaan. Het voldoende bespreken van elk orgaan neemt per model vier pagina's in. Een beknopte bespreking bevindt zich in sectie 4.4.

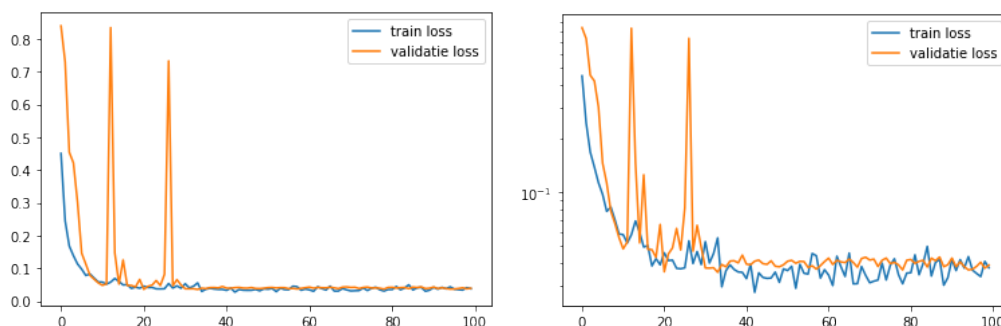
### 3.4.1 Longen

Voor de longen gebruikten we weer de configuratie van drie schijnpatiënten en werd er getraind voor 100 epochs. Opnieuw waren we, gezien het trainen op slechts 1 patiënt, niet expliciet op zoek naar een “getraind” model, maar naar een overtraind model. We willen namelijk weten of het model trainbaar is en of het klaar is om getraind te worden op meer patiënten om zo een werkend model te verkrijgen. Het overtrainen zal zich minimaal uiten in het stagneren van de metrieken ten gevolge van de data-augmentatie en zich maximaal uiten in het uiteindelijk verslechteren van de validatiemetrieken indien het model alsnog in staat is de willekeurige variaties van de trainingsdata (ten gevolge van de data-augmentatie) te leren. Als eerste (visuele) indicator kijken we naar het verloop van de IoU score van de trainset en validatieset tijdens het trainen:



Figuur 55: Verloop van de Dicescore op de trainset en de validatieset tijdens het trainen van de longen over een periode van 100 epochs. Lineaire schaal (links) en logschaal (rechts).

We zien dat het model al zeer snel zeer hoge Dicescores behaalt, zowel voor de trainpatiënt als de validatiepatiënt. Dit was te verwachten bij de longen. Voor de losswaarden hebben we de volgende grafieken:



Figuur 56: Verloop van de losswaarde op de trainset en de validatieset tijdens het trainen van de longen over een periode van 100 epochs. Lineaire schaal (links) en logschaal (rechts).

Aangezien de losswaarden direct gecorreleerd zijn met de Dice score valt er niet meteen nieuwe informatie te zien. We hebben wel een beter beeld op de fluctuaties van het model en we zien dat deze relatief volatiel zijn. Het model lijkt redelijk geconvergeerd.

De maximale train Dicescore bedroeg 0.972 terwijl de test Dicescore 0.960 bedroeg. De minimale train Diceloss bedroeg 0.027 terwijl de test Diceloss 0.101 was. Deze metrieken zijn van hetzelfde kaliber en we zouden nu al kunnen concluderen dat het model trainbaar is.

Opnieuw bekijken we de confusionmatrices om te zien waar er eventueel nog verbetering kan.

Tabel 33: Confusionmatrix van de testpatiënt na 100 epochs bij de longen.

Voorspelling/Ground Truth	Long (P = 1268382)	Niet Long(N = 6792546)
Long	True Positive (TP) = 1222168	False Positive (FP) = 18707
Niet Long	False Negative (FN) = 46214	True Negative (TN) = 6773839

Tabel 34: Confusionmatrix van de trainpatiënt na 100 epochs bij de longen.

Voorspelling/Ground Truth	Long (P = 983081)	Niet Long(N = 6094807)
Long	True Positive (TP) = 965731	False Positive (FP) = 16251
Niet Long	False Negative (FN) = 17350	True Negative (TN) = 6078556

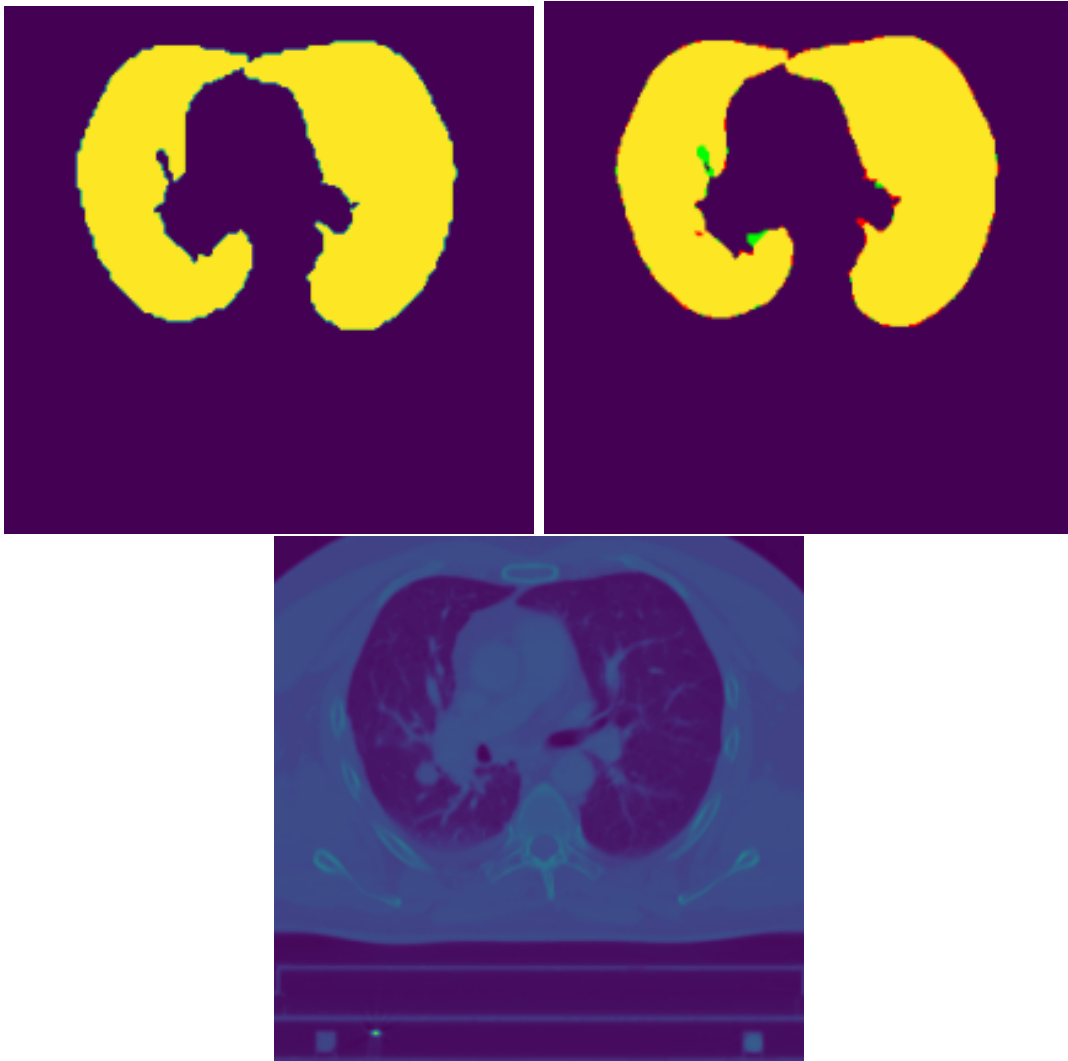
Indien we deze gegevens verwerken bekommen we volgende kwantitatieve indicatoren:

Tabel 35: True positive rate (TPR), true negative rate (TNR), false positive rate (FPR) en false negative rate (FNR) voor de test- en trainpatiënt na 100 epochs.

Patient	TPR = TP/P	TNR = TN/N	FPR = FP/N	FNR = FN/P
Test	0.964	0.997	0.003	0.036
Train	0.982	0.997	0.003	0.018

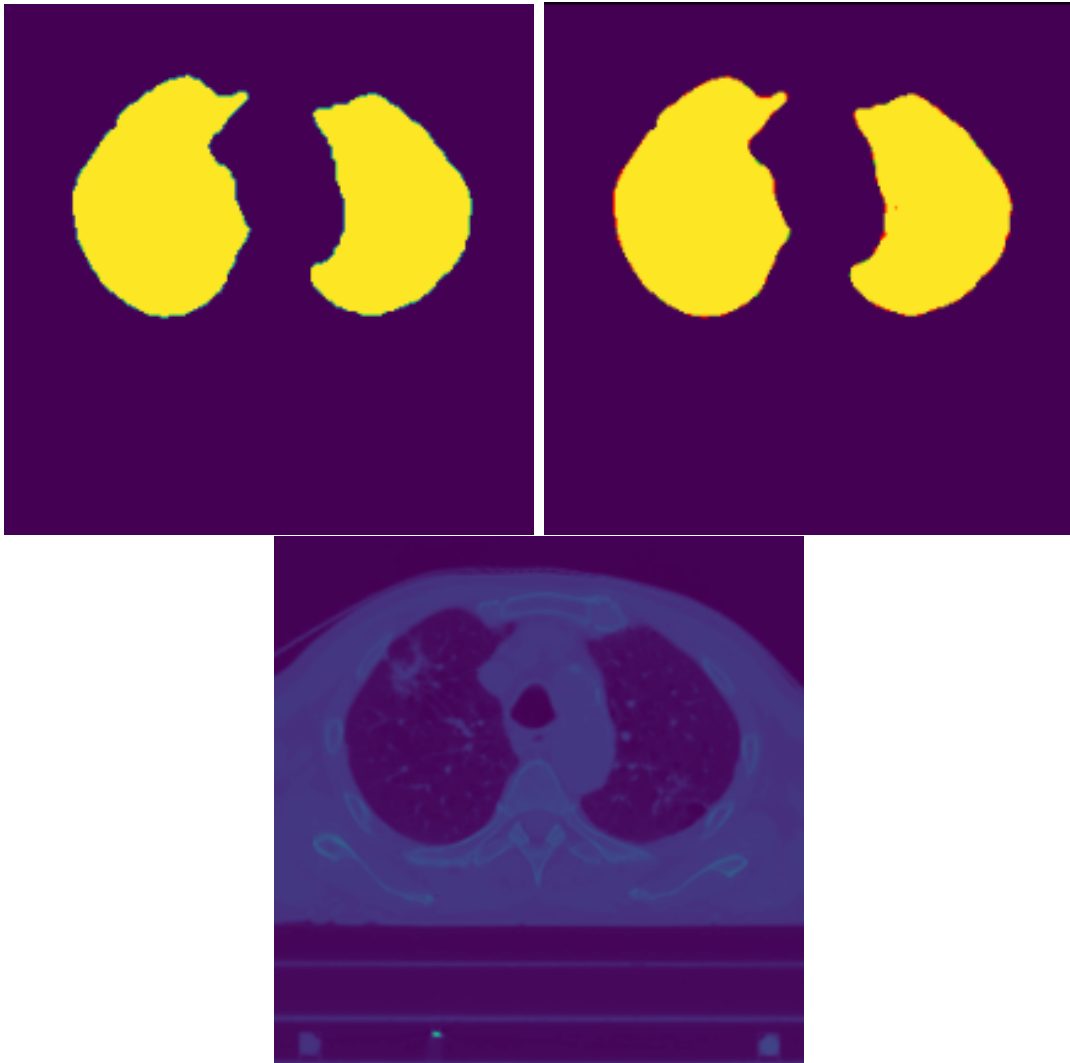
Over de grote lijn zijn deze resultaten meer dan wenselijk. We hadden dan ook geen probleem verwacht voor de longen.

We kijken verder naar visuele voorspellingen van het model om te zien waar de kleine fouten exact zitten. Paars is TN, geel is TP, groen is FP en rood is FN. De eerste set is afkomstig van de trainpatiënt en de tweede set van de testpatiënt.



Figuur 57: CT-beeld (onder) ground truth longsegmentatie (linksboven) voorspelling experiment (rechtsboven) van de trainpatiënt.

We zien dat het model bijna perfect is. Er zijn hier en daar nog kleine fouten aan grillige randen. Deze fouten zijn niet erg aangezien ze binnen de foutenmarges van de apparatuur vallen.



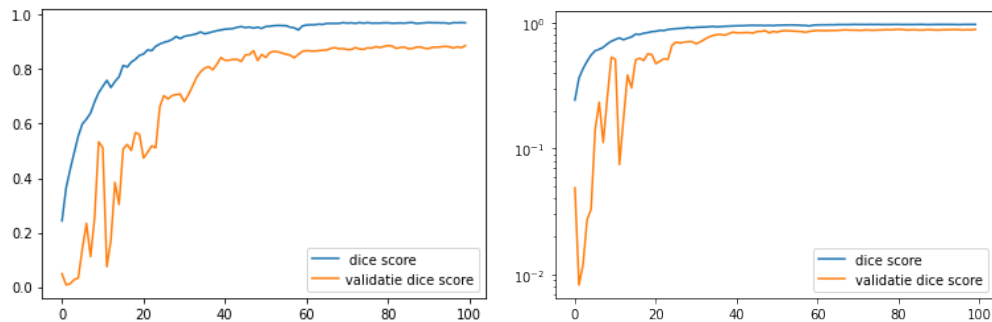
Figuur 58: CT-beeld (onder) ground truth longsegmentatie (linksboven) voorspelling experiment (rechtsboven) van de testpatiënt.

Ook bij de testpatiënt zien we uitstekende voorspellingen. Opnieuw hier en daar kleine fouten aan de randen maar niks dat een probleem kan vormen.

We kunnen hierbij concluderen dat het model trainbaar is op de longen wat een goede indicator is voor het bruikbaar zijn van het model op andere organen.

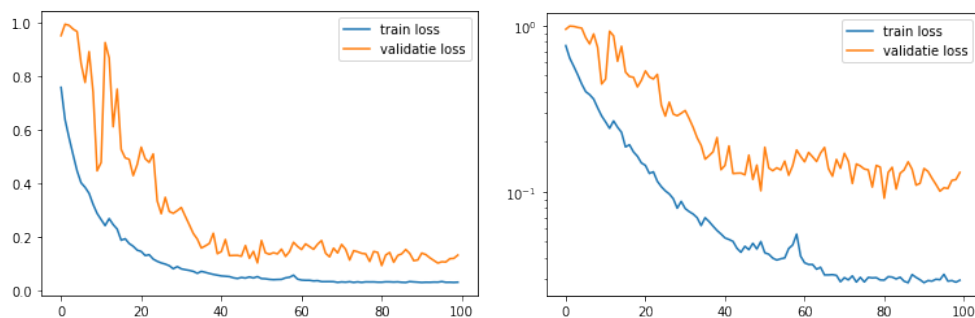
### 3.4.2 Hart

Voor het hart gebruikten we weer de configuratie van drie schijnpatiënten en werd er getraind voor 100 epochs. Opnieuw waren we, gezien het trainen op slechts 1 patiënt, niet expliciet op zoek naar een “getraind” model maar naar een overtraind model. We willen namelijk weten of het model trainbaar is en of het klaar is om getraind te worden op meer patiënten om zo een werkend model te verkrijgen. Het overtrainen zal zich minimaal uiten in het stagneren van de metrieken ten gevolge van de data-augmentatie en zich maximaal uiten in het uiteindelijk verslechteren van de validatiemetrieken indien het model alsnog in staat is de willekeurige variaties van de trainingsdata (ten gevolge van de data-augmentatie) te leren. Als eerste (visuele) indicator kijken we naar het verloop van de IoU score van de trainset en validatieset tijdens het trainen:



Figuur 59: Verloop van de Dicescore op de trainset en de validatieset tijdens het trainen van het hart over een periode van 500 epochs. Lineaire schaal (links) en logschaal (rechts).

We zien weer redelijk normale trainingscurves. Ook nu liggen de metrieken voor de trainpatiënt beduidend beter dan de metrieken voor de validatiepatiënt. Dit hadden we ook bij het gewogen binaire lossfunctie model. We zien naar het einde toe een redelijke afvlakking wat wijst op het getraind zijn van het model. We kijken verder naar de losswaarden:



Figuur 60: Verloop van de losswaarde op de trainset en de validatieset tijdens het trainen van het hart over een periode van 500 epochs. Lineaire schaal (links) en logschaal (rechts).

Opnieuw kunnen we bij de losswaarden beter het grillige karakter van het trainproces zien. Ondanks de grilligheid lijkt het model visueel te stagneren naar het einde toe waardoor we weer kunnen concluderen dat het model getraind is.



De maximale train Dicescore bedroeg 0.971 terwijl de test Dicescore 0.876 bedroeg. De minimale train Diceloss bedroeg 0.028 terwijl de test Diceloss 0.186 was. Deze metrieken zijn niet compleet van hetzelfde kaliber wat wijst op een relatief sterke overtraining of dus algemeen het (over)trainbaar zijn van het model.

We kijken verder naar de confusionmatrices om te zien wat de oorzaak is van de wat lagere test Dicescore.

Tabel 36: Confusionmatrix van de testpatiënt na 100 epochs op het hart.

Voorspelling/Ground Truth	Hart (P = 234508)	Niet Hart(N = 2518004)
Hart	True Positive (TP) = 202019	False Positive (FP) = 17751
Niet Hart	False Negative (FN) = 32489	True Negative (TN) = 2500253

Tabel 37: Confusionmatrix van de trainpatiënt na 100 epochs op het hart.

Voorspelling/Ground Truth	Hart (P = 147041)	Niet Hart(N = 1491359)
Hart	True Positive (TP) = 145172	False Positive (FP) = 1931
Niet Hart	False Negative (FN) = 1869	True Negative (TN) = 1489428

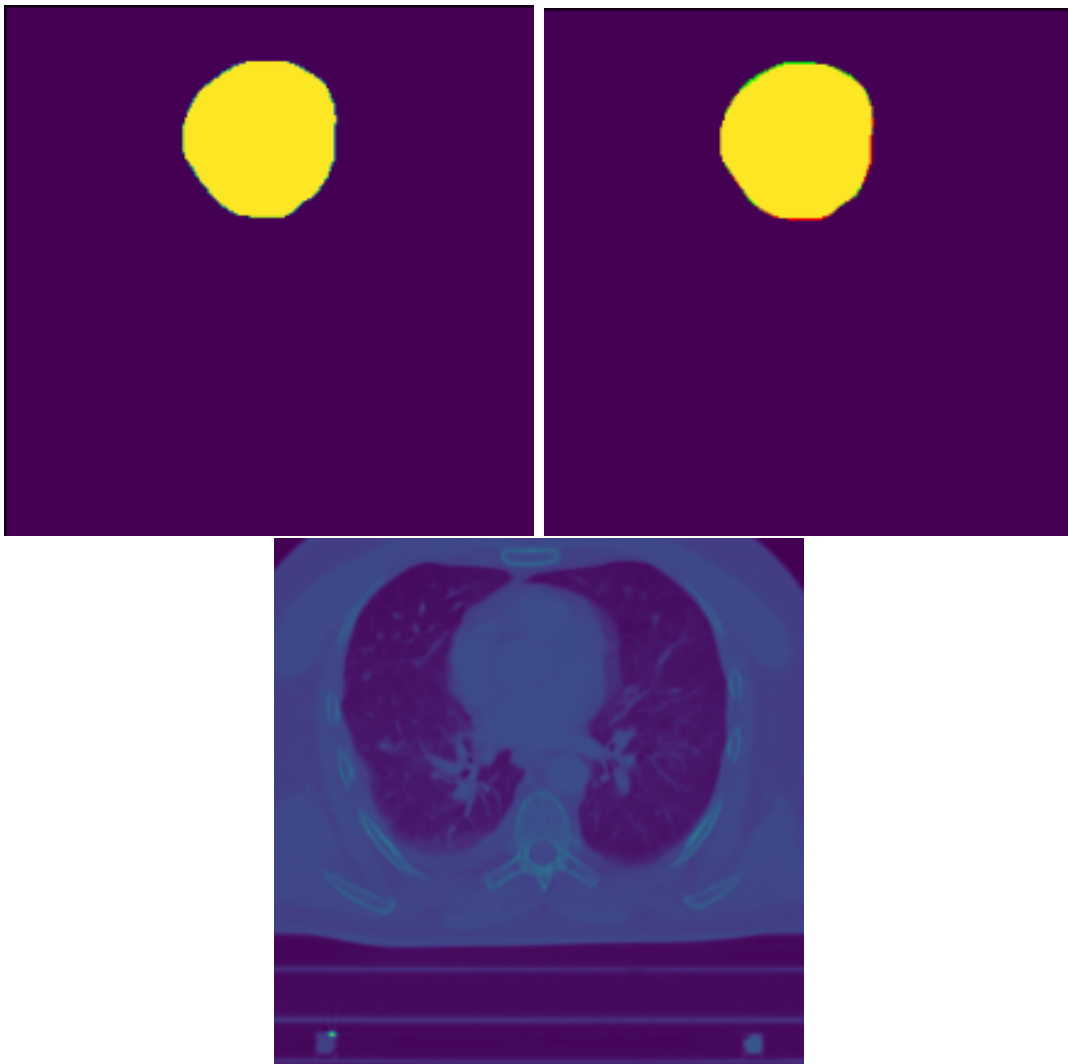
Indien we deze gegevens verwerken bekomen we volgende kwantitatieve indicatoren:

Tabel 38: True positive rate (TPR), true negative rate (TNR), false positive rate (FPR) en false negative rate (FNR) voor de test- en trainpatiënt na 500 epochs.

Patient	TPR = TP/P	TNR = TN/N	FPR = FP/N	FNR = FN/P
Test	0.861	0.993	0.007	0.139
Train	0.987	0.999	0.001	0.013

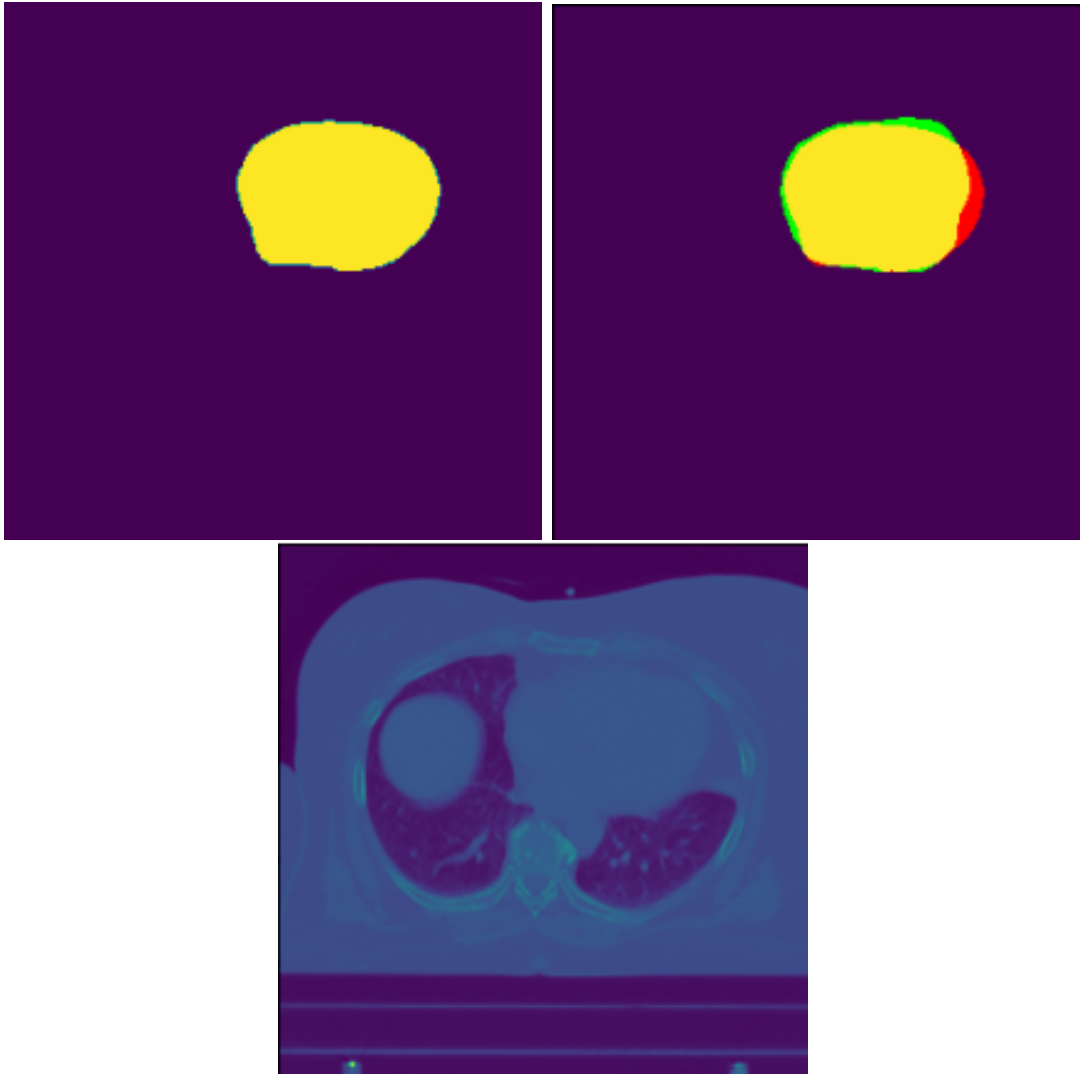
We zien dat de test TPR en zo ook FNR toch nog niet wenselijk is. Deze metrieken hebben we graag optimaler aangezien het correct annoteren van het orgaan belangrijk is voor de stralingsplanning. Men wilt namelijk niet dat een orgaan niet compleet geannoteerd is en als gevolg een meer dan toegestane dosis krijgt.

We kijken verder naar de voorspellingen om te zien of het model de organen voldoende goed kan segmenteren. Paars is TN, geel is TP, groen is FP en rood is FN. De eerste set is afkomstig van de trainpatiënt en de tweede set van de testpatiënt.



Figuur 61: CT-beeld (onder) ground truth hartsegmentatie (linksboven) voorspelling experiment (rechtsboven) van de trainpatiënt.

We zien uitstekende voorspellingen voor de trainpatiënt. Dit is een goede indicator dat het model te (over)trainen valt en we primair meer data nodig hebben.



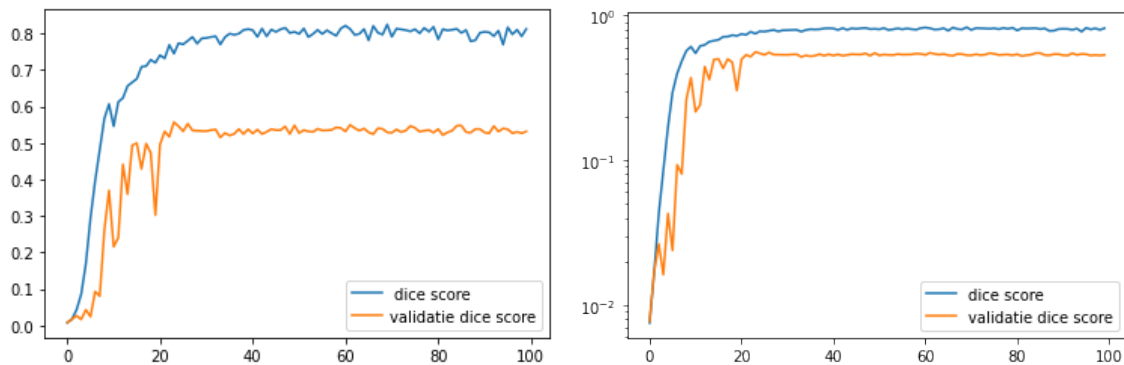
Figuur 62: CT-beeld (onder) ground truth hartsegmentatie (linksboven) voorspelling experiment (rechtsboven) van de testpatiënt.

Bij de testpatiënt zien we zoals verwacht toch nog redelijk wat valse negatieven. Algemeen lijkt de vorm wel in orde.

Men kan concluderen dat dit model te (over)trainen valt en simpelweg meer (gevarieerde) data nodig heeft om ook uitstekende resultaten te behalen op de testpatiënten.

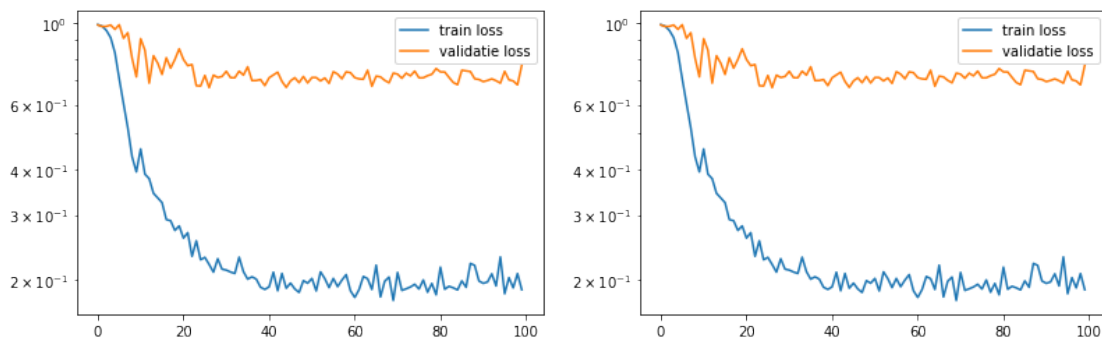
### 3.4.3 Slokdarm

Voor de slokdarm werd er weer gestart met de configuratie van drie schijnpatiënten en werd er getraind voor 100 epochs. Opnieuw waren we, gezien het trainen op slechts 1 patiënt, niet expliciet op zoek naar een “getraind” model maar naar een overtraind model. We willen namelijk weten of het model trainbaar is en of het klaar is om getraind te worden op meer patiënten om zo een werkend model te verkrijgen. Het overtrainen zal zich minimaal uiten in het stagneren van de metrieken ten gevolge van de data-augmentatie en zich maximaal uiten in het uiteindelijk verslechteren van de validatiemetrieken indien het model alsnog in staat is de willekeurige variaties van de trainingsdata (ten gevolge van de data-augmentatie) te leren. Als eerste (visuele) indicator kijken we naar het verloop van de IoU score van de trainset en validatieset tijdens het trainen:



Figuur 63: Verloop van de Dicescore op de trainset en de validatieset tijdens het trainen op de slokdarm over een periode van 100 epochs. Lineaire schaal (links) en logschaal (rechts).

We zien dat het model al zeer vroeg stagneert en dat de trainscore beduidend hoger ligt dan de validatiescore. Dit wijst op het niet voorhanden zijn van voldoende (gevarieerde) trainingsdata. Dit is echter wat we willen aangezien dit indiceert dat het model te (over)trainen valt. We zien ook hetzelfde bij de losswaarden:



Figuur 64: Verloop van de losswaarde op de trainset en de validatieset tijdens het trainen op de slokdarm over een periode van 100 epochs. Lineaire schaal (links) en logschaal (rechts).

Inderdaad: ook hier merken we de vroege stagnatie en het beduidend verschil in losswaarden op.

De maximale train Dicescore bedroeg 0.825 terwijl de test Dicescore 0.477 bedroeg. De minimale train Dicoloss bedroeg 0.175 terwijl de test Dicoloss 0.517 was. We zien weer beduidend lagere (en slechtere) testmetrieken maar redelijk goede trainmetrieken. Dit wijst weer op het (over)trainbaar zijn van het model.

We bekijken weer de confusionmatrices om af te leiden waar het model exact fouten maakt.

Tabel 39: Confusionmatrix van de testpatiënt na 100 epochs op de slokdarm.

Voorspelling/Ground Truth	Slokdarm (P = 13624)	Niet Slokdarm(N = 6998728)
Slokdarm	True Positive (TP) = 5034	False Positive (FP) = 2060
Niet Slokdarm	False Negative (FN) = 8590	True Negative (TN) = 6996668

Tabel 40: Confusionmatrix van de trainpatiënt na 100 epochs op de slokdarm.

Voorspelling/Ground Truth	Slokdarm (P = 10292)	Niet Slokdarm(N = 6739916)
Slokdarm	True Positive (TP) = 8620	False Positive (FP) = 1073
Niet Slokdarm	False Negative (FN) = 1672	True Negative (TN) = 6738843

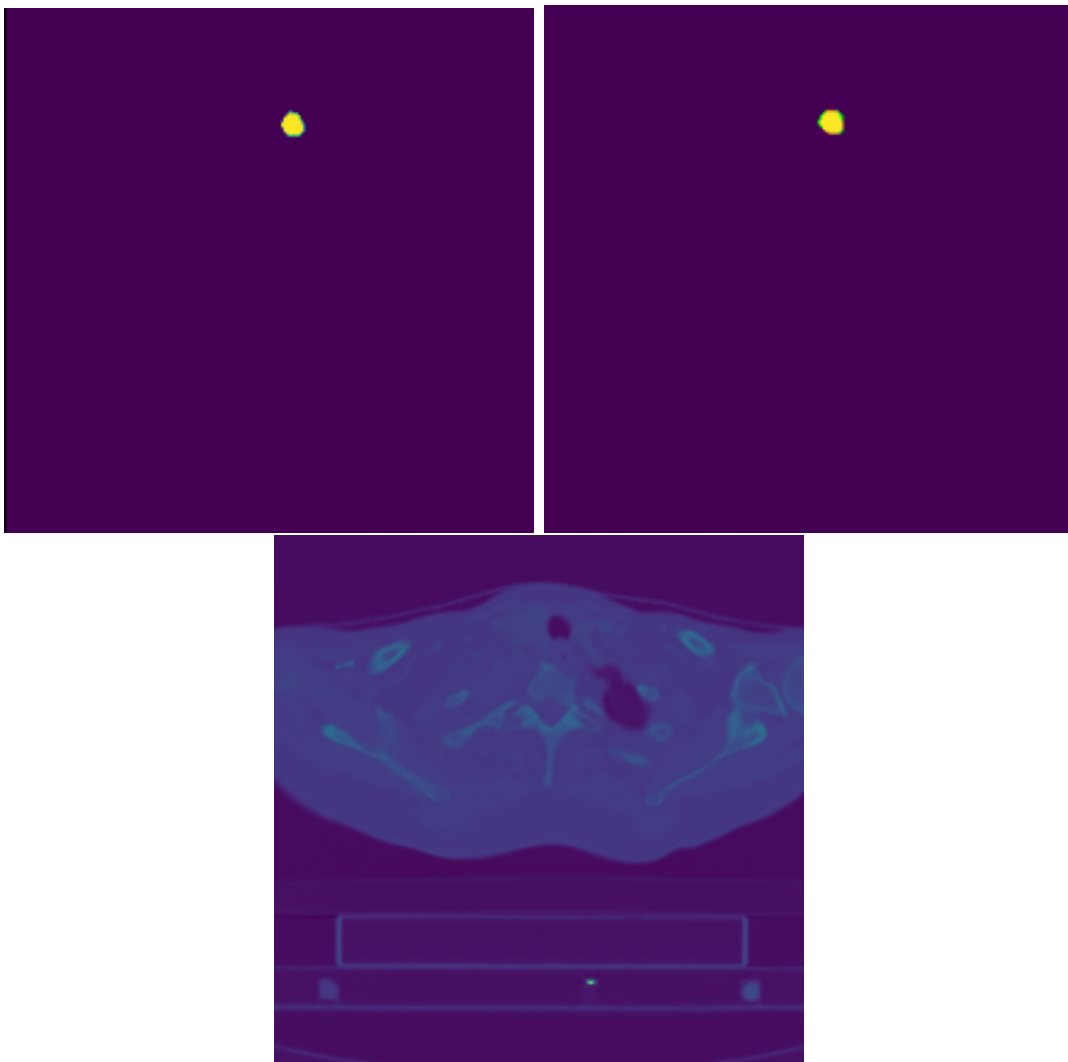
Indien we deze gegevens verwerken bekommen we volgende kwantitatieve indicatoren:

Tabel 41: True positive rate (TPR), true negative rate (TNR), false positive rate (FPR) en false negative rate (FNR) voor de test- en trainpatiënt na 100 epochs.

Patient	TPR = TP/P	TNR = TN/N	FPR = FP/N	FNR = FN/P
Test	0.369	0.9997	0.0001	0.631
Train	0.838	0.9998	0.0002	0.162

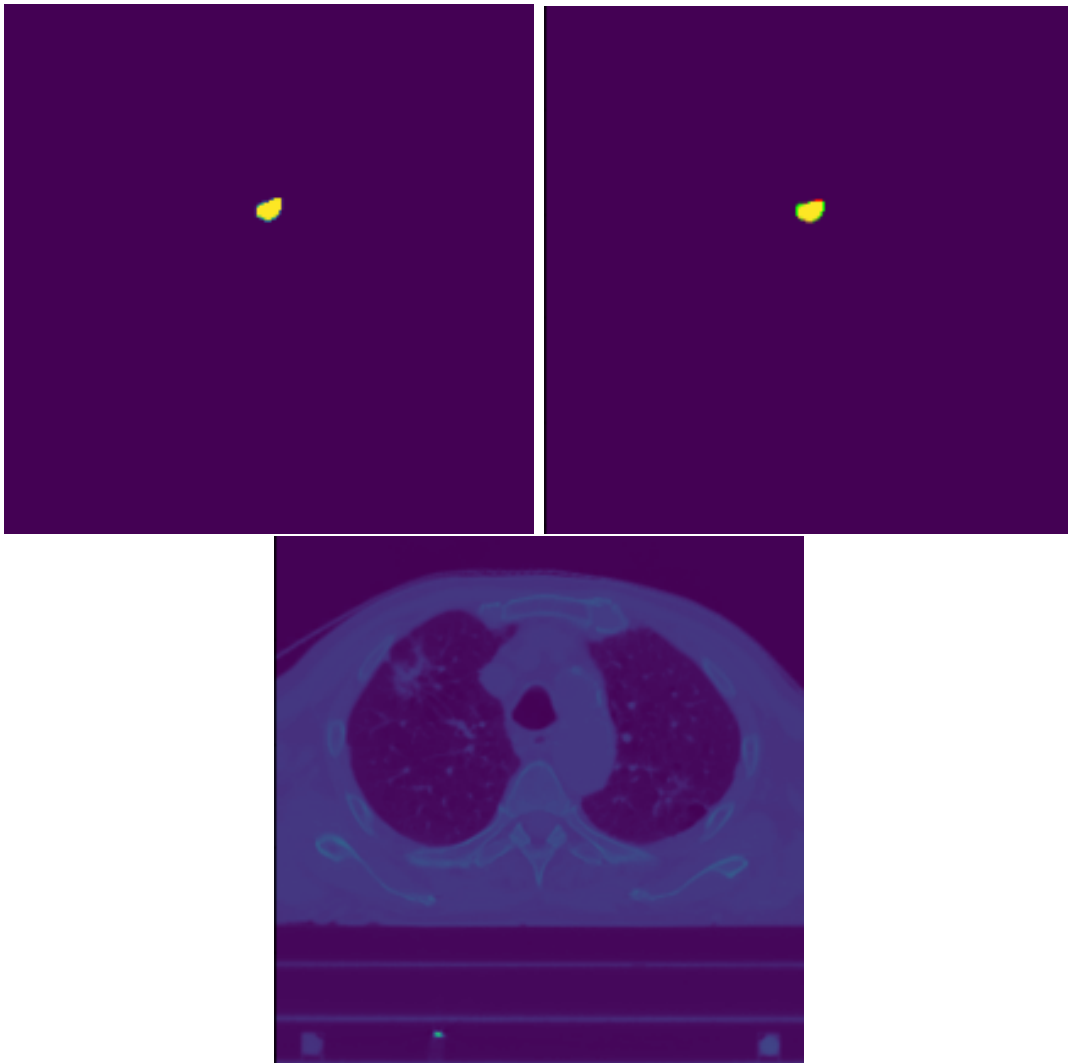
We zien een zeer lage test TPR. Echter laat ook de train TPR het afweten. Vermoedelijk waren er weinig slokdarmfoto's waardoor de data-augmentatie de trainingsdata te gevarieerd maakten. Deze problemen zijn op te lossen door meer trainingsdata, wat dus weer een indicator is van het (over)trainbaar zijn van het model.

We kijken weer naar enkele voorspellingen om te zien hoe al dan niet erg de gevolgen zijn van de slechte TPR. Paars is TN, geel is TP, groen is FP en rood is FN. De eerste set is afkomstig van de trainpatiënt en de tweede set van de testpatiënt.



Figuur 65: CT-beeld (onder) ground truth slokdarmsegmentatie (linksboven) voorspelling experiment (rechtsboven) van de trainpatiënt.

Het getekend orgaan is nog redelijk in orde in vergelijking met de voorspelling. De reden van de relatief lage TPR ligt aan het feit dat het orgaan relatief klein is en de onzekerheden aan de rand daardoor zwaar doorwegen. Als gevolg zullen deze randfluctuaties een relatief groot deel van de totale oppervlakte innemen en zo grote invloeden hebben op de TPR. Echter zijn dit soort randfluctuaties niet zo erg gezien de precisie van de apparatuur.



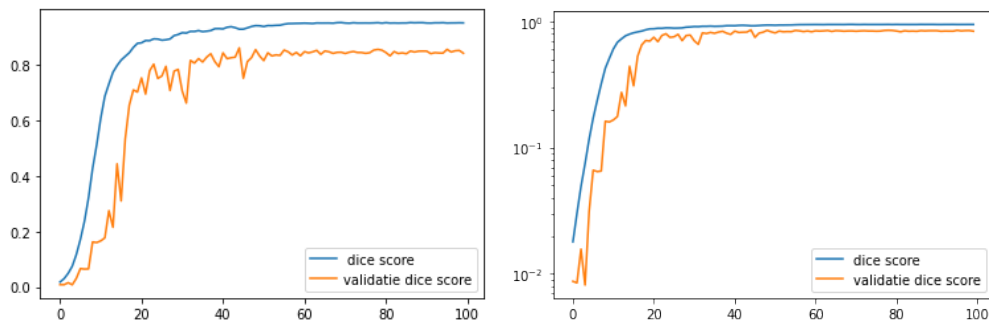
Figuur 66: CT-beeld (onder) ground truth slokdarmsegmentatie (linksboven) voorspelling experiment (rechtsboven) van de testpatiënt.

Ook bij de testpatiënt merken we hetzelfde op als bij de trainpatiënt. Fluctuaties aan de rand zorgen voor uitvergroete fouten bij de TPR. Algemeen is het echter een redelijke voorspelling.

De fouten die nog aanwezig zijn, zijn enerzijds weg te werken met meer trainingsdata en anderzijds niet van de orde dat ze relevant zijn voor het bestralingsproces. Ook dit orgaan kunnen we bij deze (over)trainbaar verklaren.

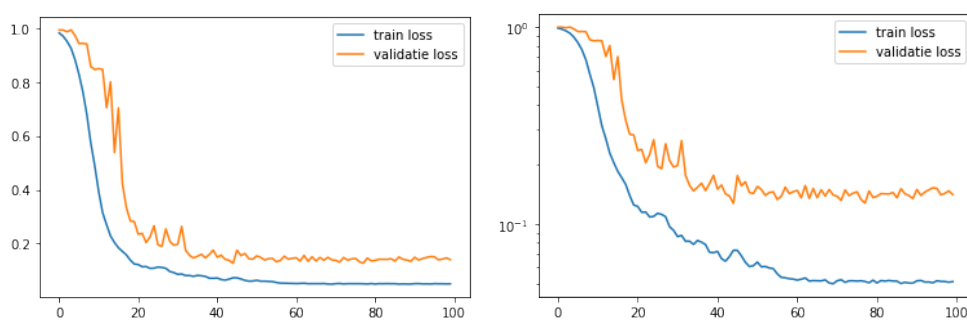
### 3.4.4 Luchtpijp

Voor de luchtpijp werd er weer gestart met de configuratie van drie schijnpatiënten en werd er getraind voor 100 epochs. Opnieuw waren we, gezien het trainen op slechts 1 patiënt, niet expliciet op zoek naar een “getraind” model maar naar een overtraind model. We willen namelijk weten of het model trainbaar is en of het klaar is om getraind te worden op meer patiënten om zo een werkend model te verkrijgen. Het overtrainen zal zich minimaal uiten in het stagneren van de metrieken ten gevolge van de data-augmentatie en zich maximaal uiten in het uiteindelijk verslechteren van de validatiemetrieken indien het model alsnog in staat is de willekeurige variaties van de trainingsdata (ten gevolge van de data-augmentatie) te leren. Als eerste (visuele) indicator kijken we naar het verloop van de IoU score van de trainset en validatieset tijdens het trainen:



Figuur 67: Verloop van de Dicescore op de trainset en de validatieset tijdens het trainen van de luchtpijp over een periode van 100 epochs. Lineaire schaal (links) en logschaal (rechts).

Net als bij de vorige organen merken we een snelle stagnatie van de metrieken op. Daarnaast zijn deze metrieken, en zeker deze van de trainpatiënt, redelijk in orde. Hetzelfde is te zien bij de losswaarden:



Figuur 68: Verloop van de losswaarde op de trainset en de validatieset tijdens het trainen van de luchtpijp over een periode van 100 epochs. Lineaire schaal (links) en logschaal (rechts).

We kunnen nu duidelijker de fluctuaties op de metrieken naar het einde toe zien. Over de grote lijn lijken ze wel gestagneerd wat wijst op het getraind zijn van het model.



De maximale train Dicescore bedroeg 0.950 terwijl de test Dicescore 0.856 bedroeg. De minimale train Diceloss bedroeg 0.050 terwijl de test Diceloss 0.128 was. We merken weer een uitstekende trainmetriek op met een net iets mindere testmetriek. Algemeen wijst dit wel weer op het trainbaar zijn van het model.

We kijken verder naar de confusionmatrices om te zien waar de fouten liggen.

Tabel 42: Confusionmatrix van de testpatiënt na 100 epochs op de luchtpijp.

Voorspelling/Ground Truth	Luchtpijp (P = 19981)	Niet Luchtpijp(N = 3781107)
Luchtpijp	True Positive (TP) = 15683	False Positive (FP) = 760
Niet Luchtpijp	False Negative (FN) = 4298	True Negative (TN) = 3780347

Tabel 43: Confusionmatrix van de trainpatiënt na 100 epochs op de luchtpijp.

Voorspelling/Ground Truth	Luchtpijp (P = 13187)	Niet Luchtpijp(N = 3591293)
Luchtpijp	True Positive (TP) = 12716	False Positive (FP) = 283
Niet Luchtpijp	False Negative (FN) = 471	True Negative (TN) = 3591010

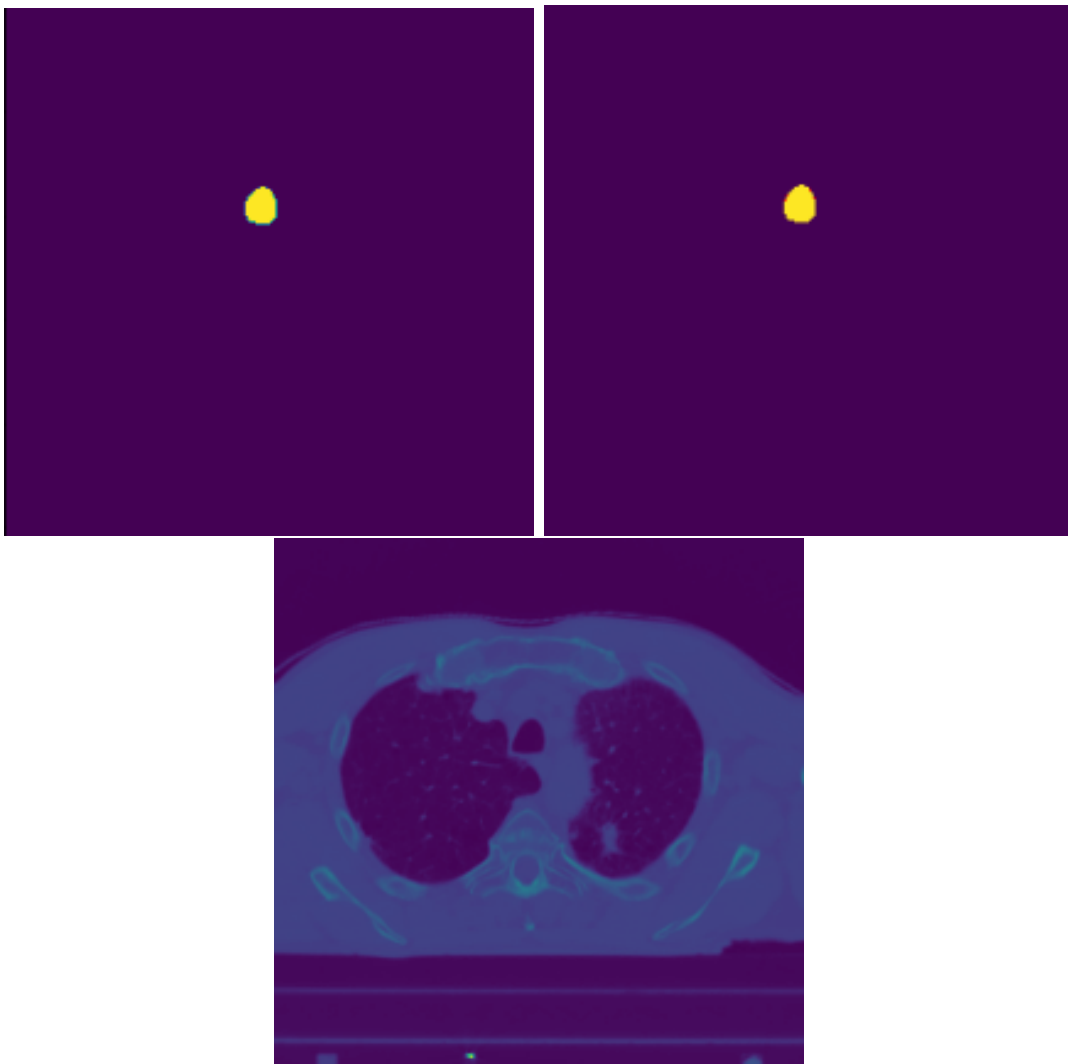
Indien we deze gegevens verwerken bekommen we volgende kwantitatieve indicatoren:

Tabel 44: True positive rate (TPR), true negative rate (TNR), false positive rate (FPR) en false negative rate (FNR) voor de test- en trainpatiënt na 100 epochs.

Patient	TPR = TP/P	TNR = TN/N	FPR = FP/N	FNR = FN/P
Test	0.785	0.9998	0.0002	0.215
Train	0.964	0.9999	0.0001	0.036

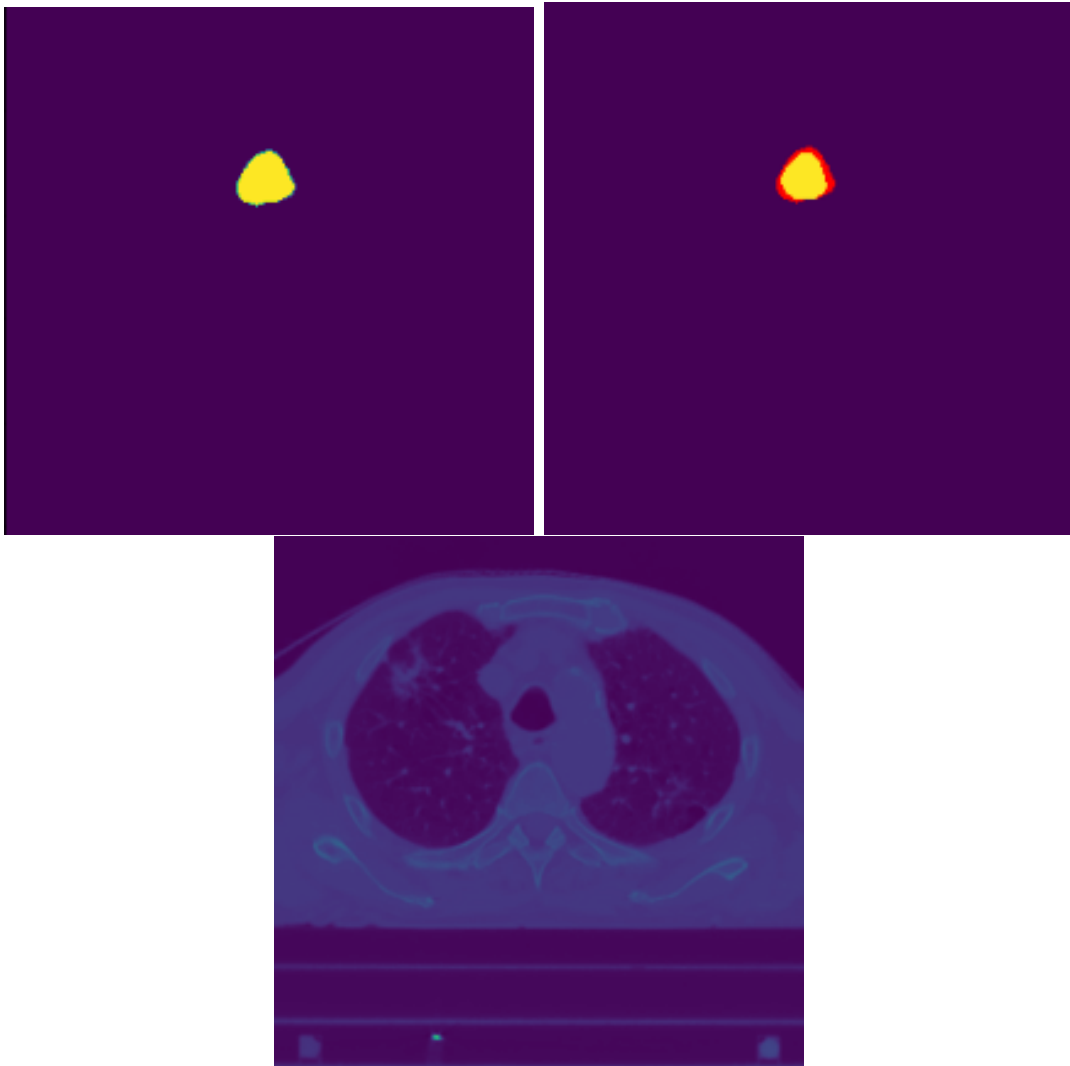
We zien weer een redelijk perfecte train TPR, maar een test TPR die wel beter moet. Vermoedelijk geldt wel weer hetzelfde probleem als het vorige orgaan waar de fouten zich vooral voortdoen aan de rand en wegens het relatief klein zijn van het orgaan dit hard wordt weer-spiegeld in de TPR.

Om te kijken of ons vermoeden correct is bekijken we weer enkele voorspellingen. Paars is TN, geel is TP, groen is FP en rood is FN. De eerste set is afkomstig van de trainpatiënt en de tweede set van de testpatiënt.



Figuur 69: CT-beeld (onder) ground truth luchtpijpsegmentatie (linksboven) voorspelling experiment (rechtsboven) van de trainpatiënt.

We zien een nagenoeg perfecte intekening van het orgaan bij de testpatiënt. Dit geeft bevestiging dat het model te trainen valt en dat er vooral meer data nodig is.



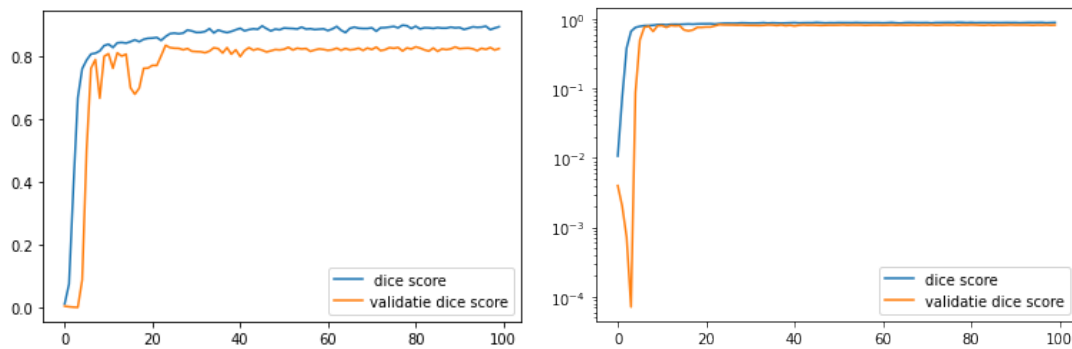
Figuur 70: CT-beeld (onder) ground truth luchtpijpsegmentatie (linksboven) voorspelling experiment (rechtsboven) van de testpatiënt.

Bij de testpatiënt zien we wel nog redelijk grote valse negatieven. Dit moet zeker beter aangezien fouten van deze orde al iets minder vergefelijk zijn en het belangrijk is het orgaan goed te annoteren om geen beschadiging van het orgaan te krijgen

De algemene conclusie is weer dat het model trainbaar is en er dus vooral meer data nodig is om even goede resultaten op de testset als op de trainset te krijgen.

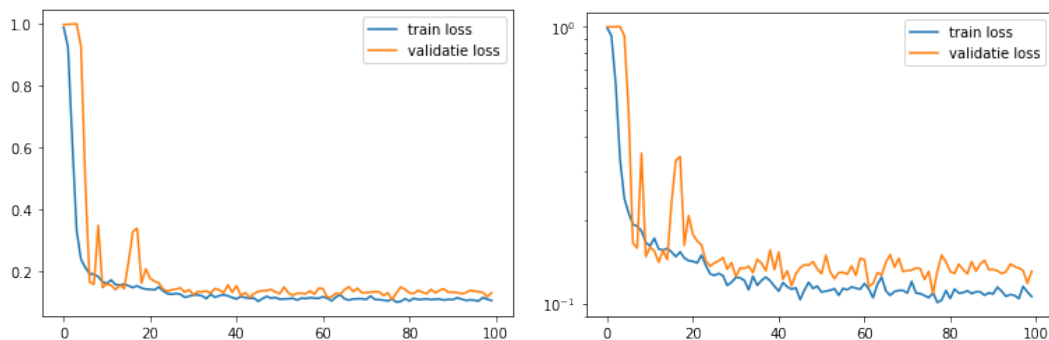
### 3.4.5 Ruggenmerg

Voor het ruggenmerg werd er weer gestart met de configuratie van drie schijnpatiënten en werd er getraind voor 100 epochs. Opnieuw waren we, gezien het trainen op slechts 1 patiënt, niet expliciet op zoek naar een “getraind” model maar naar een overtraind model. We willen namelijk weten of het model trainbaar is en of het klaar is om getraind te worden op meer patiënten om zo een werkend model te verkrijgen. Het overtrainen zal zich minimaal uiten in het stagneren van de metrieken ten gevolge van de data-augmentatie en zich maximaal uiten in het uiteindelijk verslechteren van de validatiemetrieken indien het model alsnog in staat is de willekeurige variaties van de trainingsdata (ten gevolge van de data-augmentatie) te leren. Als eerste (visuele) indicator kijken we naar het verloop van de IoU score van de trainset en validatieset tijdens het trainen:



Figuur 71: Verloop van de Dicescore op de trainset en de validatieset tijdens het trainen van het ruggenmerg over een periode van 100 epochs. Lineaire schaal (links) en logschaal (rechts).

We zien weer een vroege stagnatie van de metrieken met consistent hogere trainscore dan validatiescore. Dit geeft weer een goede bevestiging dat het model te (over)trainen valt. Om specifieker naar de fluctuaties te kijken, bekijken we ook de losswaarden:



Figuur 72: Verloop van de losswaarde op de trainset en de validatieset tijdens het trainen van het ruggenmerg over een periode van 100 epochs. Lineaire schaal (links) en logschaal (rechts).

Hier zien we hetzelfde effect maar dan met meer detail. Weer een globale stagnatie met zekere fluctuatie. Een betere bevestiging dat het model getraind is en de fluctuatie afkomstig is van de data-augmentatie.

De maximale train Dicescore bedroeg 0.899 terwijl de test Dicescore 0.798 bedroeg. De minimale train Dicoloss bedroeg 0.101 terwijl de test Dicoloss 0.210 was. We kunnen weer dezelfde opmerking maken dat het model, gezien de metrieken, in staat is te (over)trainen.

We kijken verder naar de confusionmatrices om te type fouten te bestuderen.

Tabel 45: Confusionmatrix van de testpatiënt na 100 epochs op het ruggenmerg.

Voorspelling/Ground Truth	ruggenmerg (P = 20014)	Niet ruggenmerg(N = 14987730)
ruggenmerg	True Positive (TP) = 14844	False Positive (FP) = 2009
Niet ruggenmerg	False Negative (FN) = 5170	True Negative (TN) = 14985721

Tabel 46: Confusionmatrix van de trainpatiënt na 100 epochs op het ruggenmerg.

Voorspelling/Ground Truth	ruggenmerg (P = 15602)	Niet ruggenmerg(N = 14598926)
ruggenmerg	True Positive (TP) = 14638	False Positive (FP) = 1146
Niet ruggenmerg	False Negative (FN) = 964	True Negative (TN) = 14597780

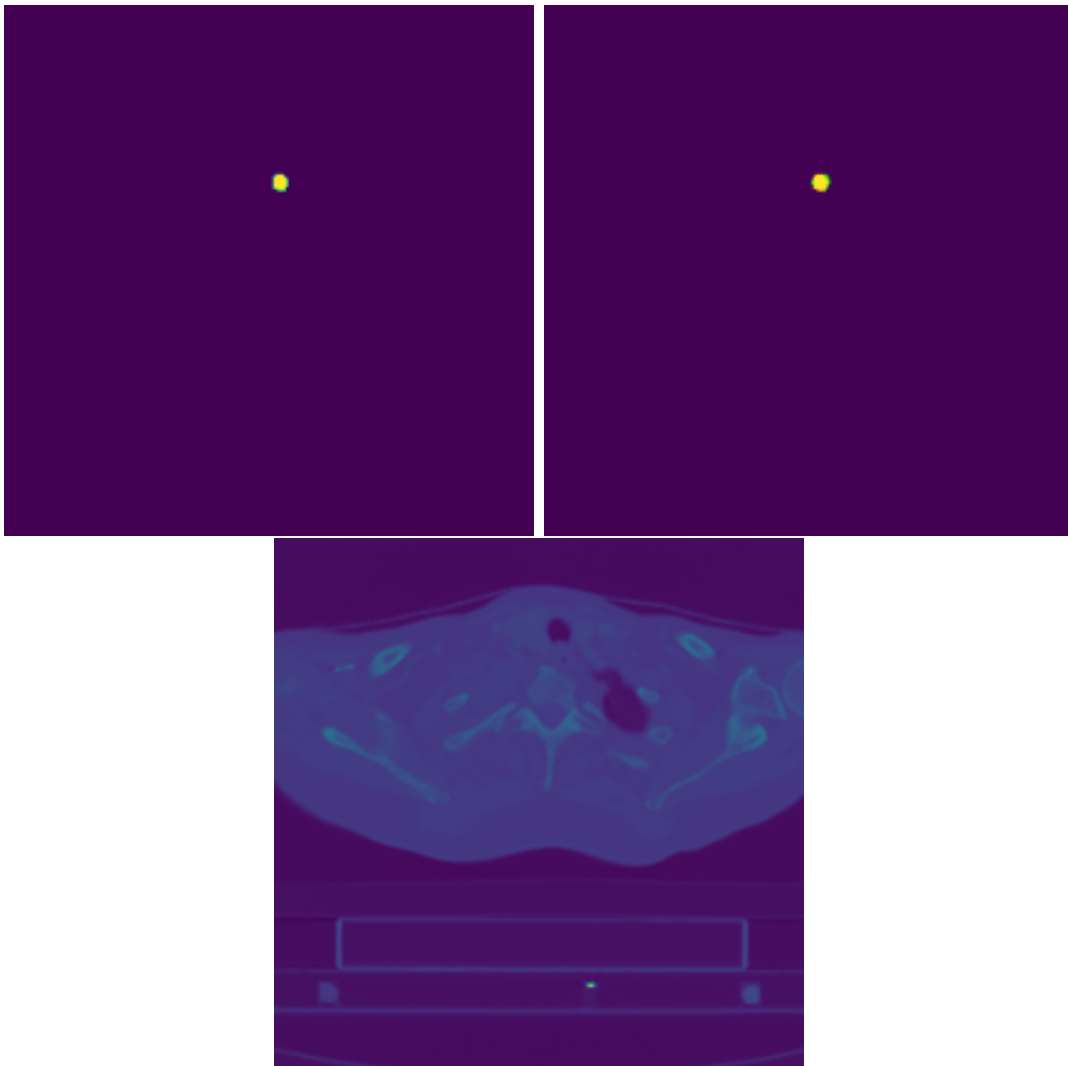
Indien we deze gegevens verwerken bekommen we volgende kwantitatieve indicatoren:

Tabel 47: True positive rate (TPR), true negative rate (TNR), false positive rate (FPR) en false negative rate (FNR) voor de test- en trainpatiënt na 100 epochs.

Patient	TPR = TP/P	TNR = TN/N	FPR = FP/N	FNR = FN/P
Test	0.742	0.9998	0.0002	0.258
Train	0.938	0.9999	0.0001	0.062

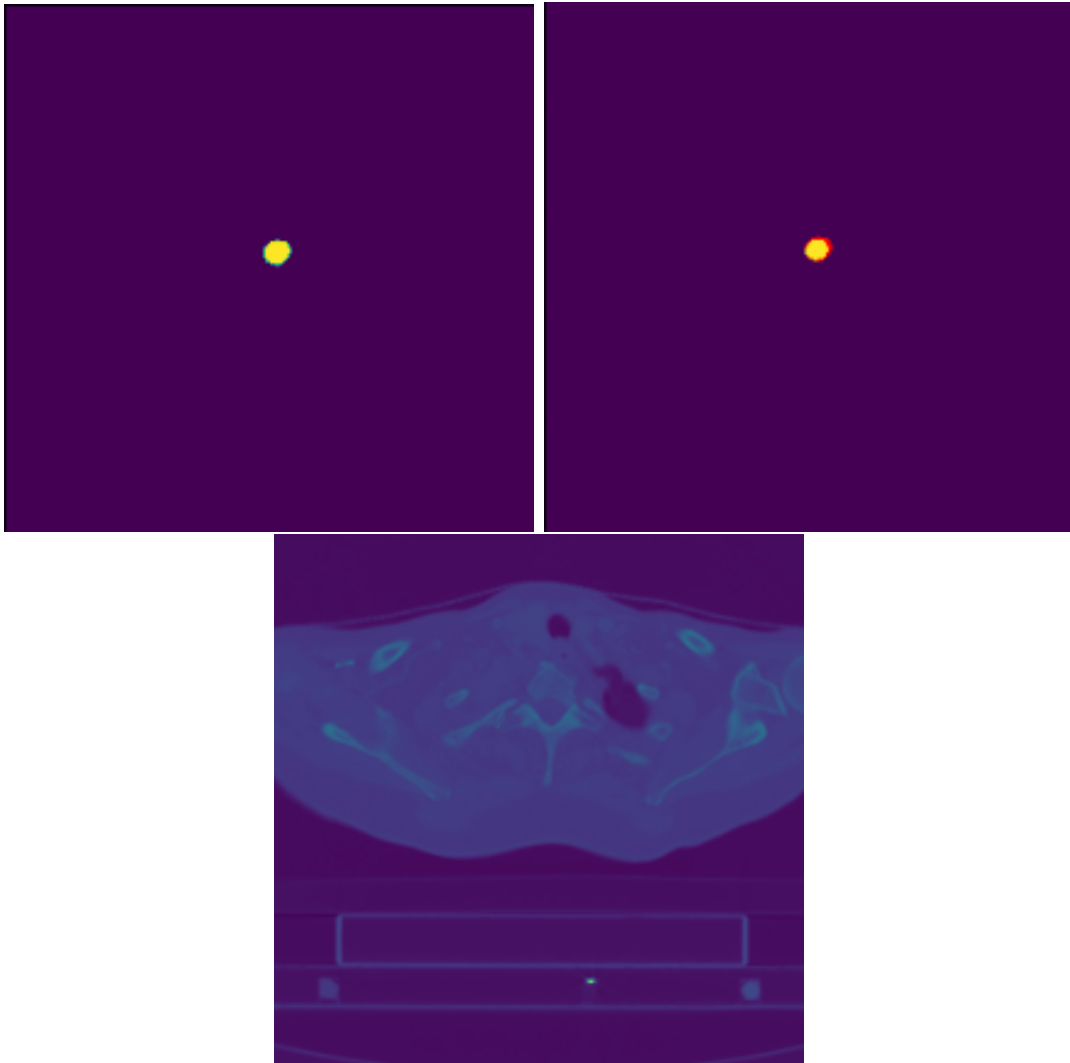
Opnieuw hebben we een redelijke train TPR maar een test TPR die beter moet. Ondertussen weten we al de verklaring van dit soort fouten bij dit type organen. Het zullen de fluctuaties aan de rand zijn die bij kleine organen zwaarder doorwegen dan bij grote organen.

We controleren weer ons vermoeden van de oorzaak van de TPR via visuele voorspellingen. Paars is TN, geel is TP, groen is FP en rood is FN. De eerste set is afkomstig van de trainpatiënt en de tweede set van de testpatiënt.



Figuur 73: CT-beeld (onder) ground truth ruggenmergsegmentatie (linksboven) voorspelling experiment (rechtsboven) van de trainpatiënt.

Bij de trainpatiënt zien we een vrij goede segmentatie voor zelfs zo een klein orgaan. Er zijn slechts minimale fouten aan de rand die natuurlijk een zekere doorweging hebben op de TPR.



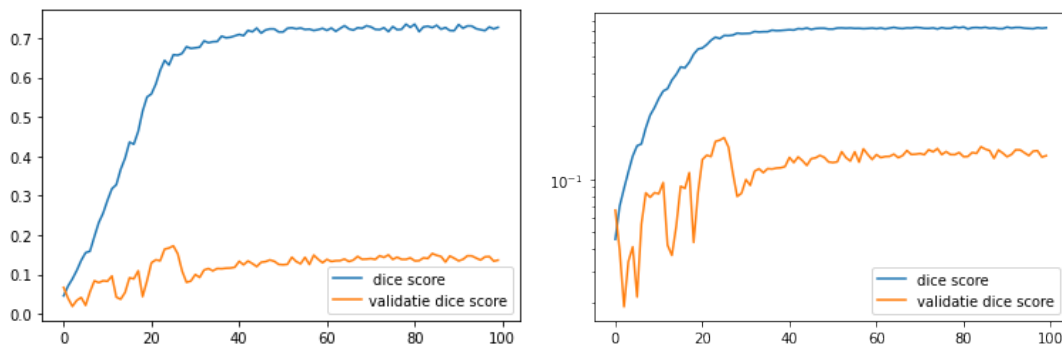
Figuur 74: CT-beeld (onder) ground truth ruggenmergsegmentatie (linksboven) voorspelling experiment (rechtsboven) van de testpatiënt.

We merken weer een relatief grote FPR op aan de rand in vergelijking met het totale orgaan. Dit moet zeker beter aangezien een te grote dosis bij het ruggenmerg een falen van de rest van het ruggenmerg kan betekenen.

Dezelfde conclusies worden getrokken. Namelijk: het model is (over)trainbaar voor dit orgaan en meer trainingsdata is nodig om verdere verbeteringen te hebben.

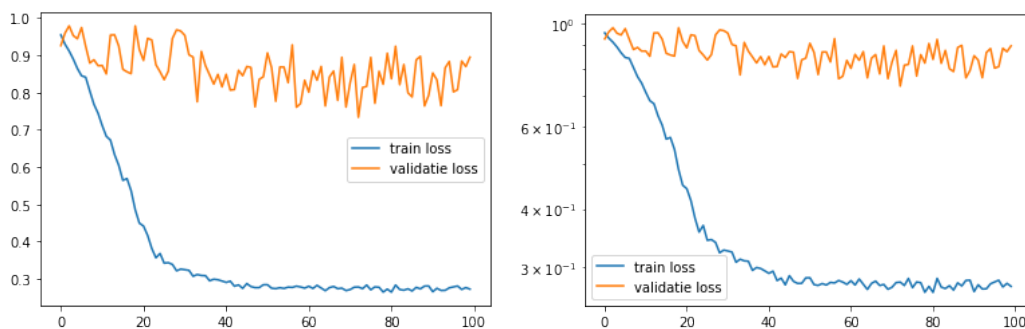
### 3.4.6 GTV

Voor het GTV werd er weer gestart met de configuratie van drie schijnpatiënten en werd er getraind voor 100 epochs. Opnieuw waren we, gezien het trainen op slechts 1 patiënt, niet expliciet op zoek naar een “getraind” model maar naar een overtraind model. We willen namelijk weten of het model trainbaar is en of het klaar is om getraind te worden op meer patiënten om zo een werkend model te verkrijgen. Het overtrainen zal zich minimaal uiten in het stagneren van de metrieken ten gevolge van de data-augmentatie en zich maximaal uiten in het uiteindelijk verslechteren van de validatiemetrieken indien het model alsnog in staat is de willekeurige variaties van de trainingsdata (ten gevolge van de data-augmentatie) te leren. Als eerste (visuele) indicator kijken we naar het verloop van de IoU score van de trainset en validatieset tijdens het trainen:



Figuur 75: Verloop van de Dicescore op de trainset en de validatieset tijdens het trainen van het GTV over een periode van 100 epochs. Lineaire schaal (links) en logschaal (rechts).

We zien een snelle stagnatie van de score, waarbij die voor de trainpatiënt redelijk hoog ligt. Ondanks het zeer gevarieerd zijn van de tumor kunnen we dus wel hoge scores bereiken. De validatiescore is echter zeer laag aangezien er grote verschillen zitten tussen de verschillende tumoren. Voor de losswaarden hebben we hetzelfde.



Figuur 76: Verloop van de losswaarde op de trainset en de validatieset tijdens het trainen van het GTV over een periode van 100 epochs. Lineaire schaal (links) en logschaal (rechts).

Ook hier zien we dus het vermoedelijk trainbaar zijn van het GTV. We mogen echter niet vergeten dat gezien de kleine hoeveelheid data het sterk mogelijk is dat het model zeer sterk overtraind is en niet per sé generaliseerbaar.



De maximale train Dicescore bedroeg 0.735 terwijl de test Dicescore 0.199 bedroeg. De minimale train Dicoloss bedroeg 0.265 terwijl de test Dicoloss 0.865 was. Dit wijst op een zeer sterke overtraining wat te verwachten valt gezien er weinig foto's met GTV zijn en deze tussen verschillende type tumoren zeer gevarieerd kunnen zijn.

We zijn geïnteresseerd in welke fouten er nog zijn bij de trainpatiënt gezien het precies segmenteren van het GTV van zeer groot belang is.

Tabel 48: Confusionmatrix van de testpatiënt na 100 epochs op het GTV.

Voorspelling/Ground Truth	GTV (P = 4768)	Niet GTV(N = 978272)
GTV	True Positive (TP) = 1374	False Positive (FP) = 3409
Niet GTV	False Negative (FN) = 3394	True Negative (TN) = 974863

Tabel 49: Confusionmatrix van de trainpatiënt na 100 epochs op het GTV.

Voorspelling/Ground Truth	GTV (P = 26261)	Niet GTV(N = 1612139)
GTV	True Positive (TP) = 25123	False Positive (FP) = 1697
Niet GTV	False Negative (FN) = 1138	True Negative (TN) = 1610442

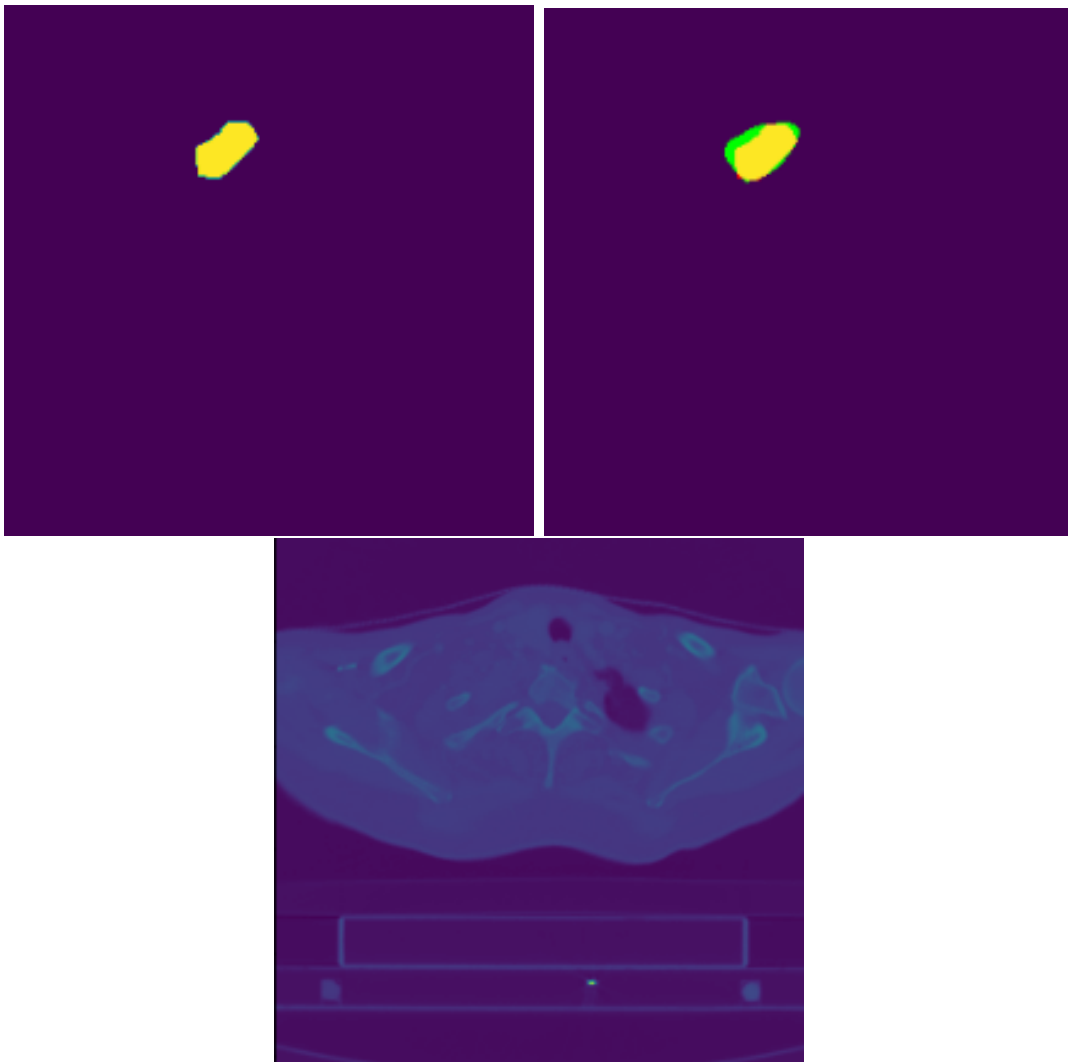
Indien we deze gegevens verwerken bekomen we volgende kwantitatieve indicatoren:

Tabel 50: True positive rate (TPR), true negative rate (TNR), false positive rate (FPR) en false negative rate (FNR) voor de test- en trainpatiënt na 100 epochs.

Patient	TPR = TP/P	TNR = TN/N	FPR = FP/N	FNR = FN/P
Test	0.288	0.997	0.003	0.712
Train	0.957	0.999	0.001	0.043

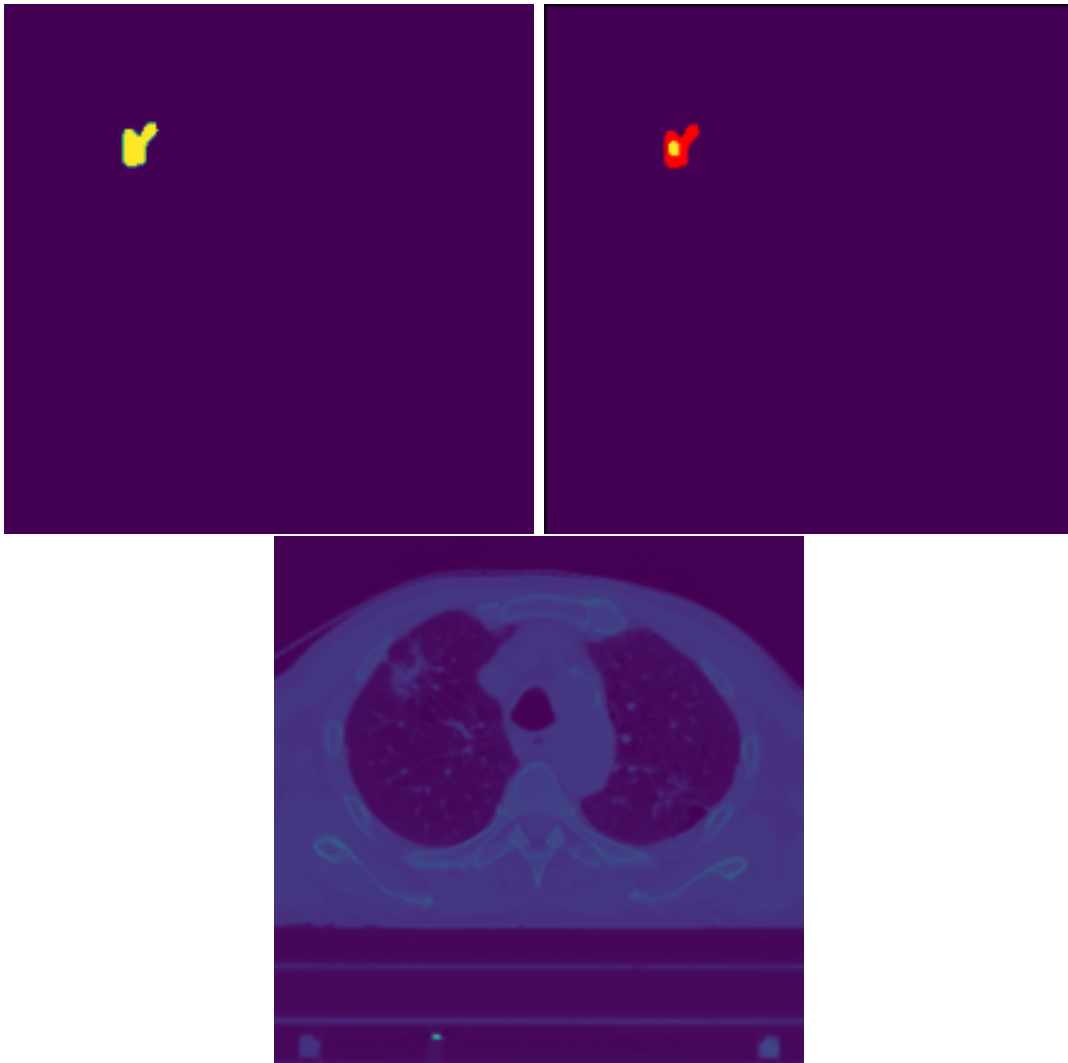
Zoals verwacht hebben we zeer slechte TPR en NR bij de testpatiënt wat wijst op het totaal niet generaliseerbaar zijn in dit experiment. De TNR van de trainpatiënt is zeer hoog wat gewenst is aangezien we zo weinig mogelijk valse positieven willen hebben. Deze valse positieven zijn namelijk sterk bepalend voor de locatie van en de hoeveelheid straling en men wilt geen onnodige (be)straling hebben. De TPR is ook redelijk hoog wat ook belangrijk is. We willen namelijk zoveel mogelijk volume van het GTV bestralen.

Om een beter beeld te krijgen op de type fouten kijken we naar enkele voorspellingen. Paars is TN, geel is TP, groen is FP en rood is FN. De eerste set is afkomstig van de trainpatiënt en de tweede set van de testpatiënt.



Figuur 77: CT-beeld (onder) ground truth GTV-segmentatie (linksboven) en voorspelling experiment (rechtsboven) van de trainpatiënt.

Zoals verwacht is de segmentatie bij de trainpatiënt redelijk in orde voor de moeilijkheidsgraad van het GTV. We zien wel nog een relatief grote regio van FP. Het is belangrijk dat deze fout niet te groot wordt zodat er niet te veel onnodig weefsel wordt bestraald.



Figuur 78: CT-beeld (onder) ground truth GTV-segmentatie (linksboven) en voorspelling experiment (rechtsboven) van de testpatiënt.

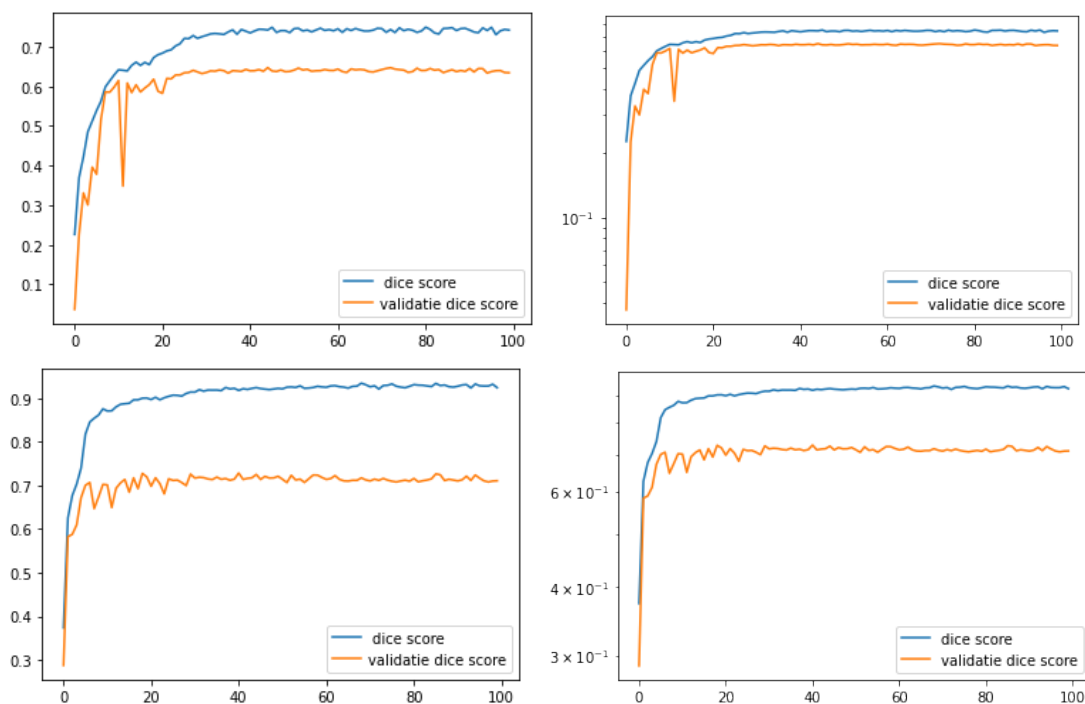
Niet wonderbaarlijk is het model zeer slecht in generaliseren en slaagt het amper in tumoren herkennen in ongeziene patiënten. Dit is problematisch aangezien een te kleine bestraalde regio niet in staat zal zijn de tumor voldoende te beschadigen, of in radiotherapie terminologie: er is onvoldoende “coverage”. Daarnaast is het model wel in staat een kleine regio aan te duiden en kan het dus anomalieën opmerken.

Het model was deze keer wel in staat de tumoren te leren herkennen bij de trainpatiënt. Echter blijft het zeer zwak in generaliseren. Hier zou het vergroten van de dataset ons veel leren of het model al dan niet mogelijk is te generaliseren.

### 3.5 Dixeloss op alle organen

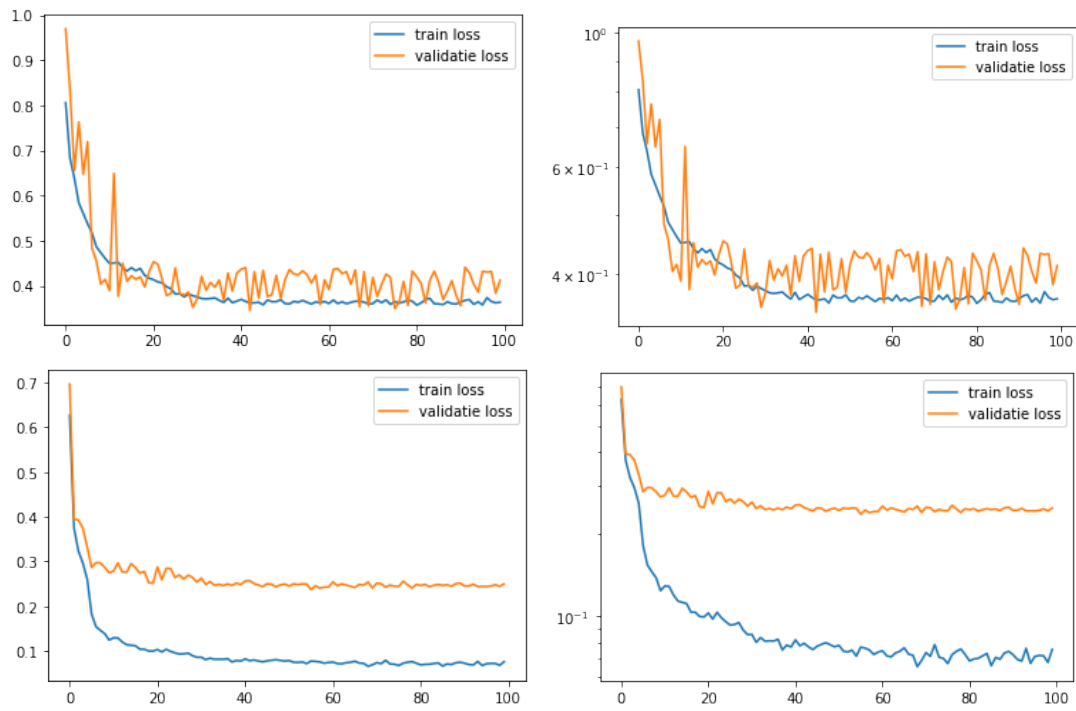
Hier werden alle organen tegelijk getraind met de Dixeloss zoals beschreven in sectie 2.2.1 over een periode van 100 epochs. Eerst op de configuratie met drie patiënten en daarna op de configuratie met 29 patiënten. Als metriek gebruiken we de Dicescore zoals beschreven in sectie 2.4.1. We zijn met deze test vooral geïnteresseerd in de verschillen tussen het trainen op de kleine groep en het trainen op de grote groep. De testen per orgaan hebben namelijk al hun potentie bewezen. Achteraf zullen we het model getraind op 29 patiënten ook testen aan de DVH en clinical goals zoals besproken in sectie 2.4.3.

We zullen eerst de traingrafieken, gewogen Dicescore en gewogen Dixeloss tussen configuratie 1 en configuratie 2 vergelijken waarbij we ons telkens beperken tot dezelfde train- en testpatiënt. Daarna zullen we voor elk orgaan apart de Dicescore, Dixeloss en voorspellingen vergelijken waarbij we weer dezelfde train- en testpatiënt gebruiken. Een beknopte bespreking bevindt zich in sectie 4.5.



Figuur 79: Verloop van de IoU score op de trainset en de validatieset tijdens het trainen van alle organen over een periode van 100 epochs. Lineaire schaal (links) en logschaal (rechts). Configuratie 1 (boven) en configuratie 2 (onder).

We zien dat configuratie 2 vlugger zijn maximum bereikt en dit ook hoger ligt. Dit is exact wat we verwachten van de groep met meer data. De modellen stagneren ook (mits een kleine fluctuatie) wat duidt op het trainbaar zijn van het model. We kijken verder naar de losswaarden:



Figuur 80: Verloop van de losswaarde op de trainset en de validatieset tijdens het trainen van alle organen over een periode van 100 epochs. Lineaire schaal (links) en logschaal (rechts). Configuratie 1 (boven) en configuratie 2 (onder).

De grafieken van de losswaarden zijn zeer gelijkend. Bij configuratie 1 zijn de fluctuaties iets meer volatiel wegens de kleinere veralgemenisering door de kleinere hoeveelheid trainingsdata. Opvallend is dat de kloof tussen test en validatie groter is bij configuratie 2 dan bij configuratie 1. Dit is vreemd aangezien we verwachten dat configuratie 2 (met meer data) beter zou zijn in het veralgemeniseren.

Tabel 51: Dicescores van de testgroep van configuratie 1 en configuratie 2.

Metriek	conf1	conf2
test Dice	0.616	0.677
test loss	0.380	0.357
max train Dice	0.750	0.935
min train loss	0.358	0.065

We zien een verbetering in de maximale train Dice van configuratie 1 naar configuratie 2. Dit is wenselijk aangezien dit betekent dat het model schaalbaar is naar grotere datasets. Ook de test Dice is verbeterd wat wijst op het voordeel van een grotere dataset in het veralgemeniseren van het model.

Voor de Dicescores werd er voordien geen threshold gebruikt. Dit kan een vertekend beeld geven. Indien een pixel 0.99 kans heeft dat het een stuk long kan zijn en 0.4 kans heeft dat het een stuk hart kan zijn, is het logisch dat we dit aanschouwen als long. De 0.4 kans voor hart geeft hierbij dan een onterechte straf op de totale score. Met dit in het achterhoofd hebben we voor dezelfde testpatiënt bij configuratie 1 en 2 de Dicescores berekend voor enkele thresholds. De eerst threshold respresenteert het model dat zekerder is van “wel dit orgaan” dan “niet dit orgaan”. De overige drie thresholds zijn arbitrair gekozen. Hiermee proberen we de zekerheid van de classificaties van het model duidelijk te maken.

Tabel 52: Dicescores met verschillende thresholds per orgaan van de testgroep van configuratie 1 en configuratie 2.

Orgaan	conf1 ( $p > 0.5$ )	conf2 ( $p > 0.5$ )	conf1 ( $p > 0.9$ )	conf2 ( $p > 0.9$ )
Hart	0.815	0.853	0.809	0.848
Slokdarm	0.495	0.630	0.468	0.619
GTV	0.031	0.086	0.028	0.078
Longen	0.968	0.976	0.963	0.973
Luchtpijp	0.772	0.808	0.754	0.804
Ruggenmerg	0.744	0.784	0.731	0.781
Orgaan	conf1 ( $p > 0.99$ )	conf2 ( $p > 0.99$ )	conf1 ( $p > 0.999$ )	conf2 ( $p > 0.999$ )
Hart	0.796	0.841	0.770	0.833
Slokdarm	0.434	0.607	0.398	0.593
GTV	0.026	0.068	0.023	0.059
Longen	0.953	0.970	0.937	0.965
Luchtpijp	0.732	0.798	0.706	0.792
Ruggenmerg	0.714	0.775	0.693	0.769

We zien dat alle organen zijn verbeterd door de grotere dataset. Ook zien we dat de Dice scores redelijk dezelfde blijven voor grotere thresholds. Dit wijst op een hoge zekerheid van het model.

We berekenen ook de IoU en Dicescores met  $p > 0.5$  voor de volledige testset per orgaan zodat ze vergeleken kunnen worden met resultaten van gelijkaardige onderzoeken.

Tabel 53: IoU en Dicescore van elke categorie van de testgroep voor configuratie 2.

Orgaan	IoU score	Dicescore
Hart	0.775	0.873
Slokdarm	0.459	0.628
GTV	0.041	0.079
Longen	0.953	0.976
luchtpijp	0.708	0.829
Ruggenmerg	0.644	0.783

Tabel 54: Herhaling tabel behaalde resultaten van gelijkaardige onderzoeken.

Orgaan	Onderzoek 1 (IoU)	Onderzoek 2 (Dice)	Onderzoek 3 (Dice)
Hart	0.817	0.941	0.85
Slokdarm	0.107	0.858	0.71
Longen	0.903	N/A	0.965
Luchtpijp	N/A	0.926	N/A
Ruggenmerg	N/A	N/A	0.83

Tabel 55: Herhaling tabel Verschillen in ingetekende contouren tussen verschillende clinici.

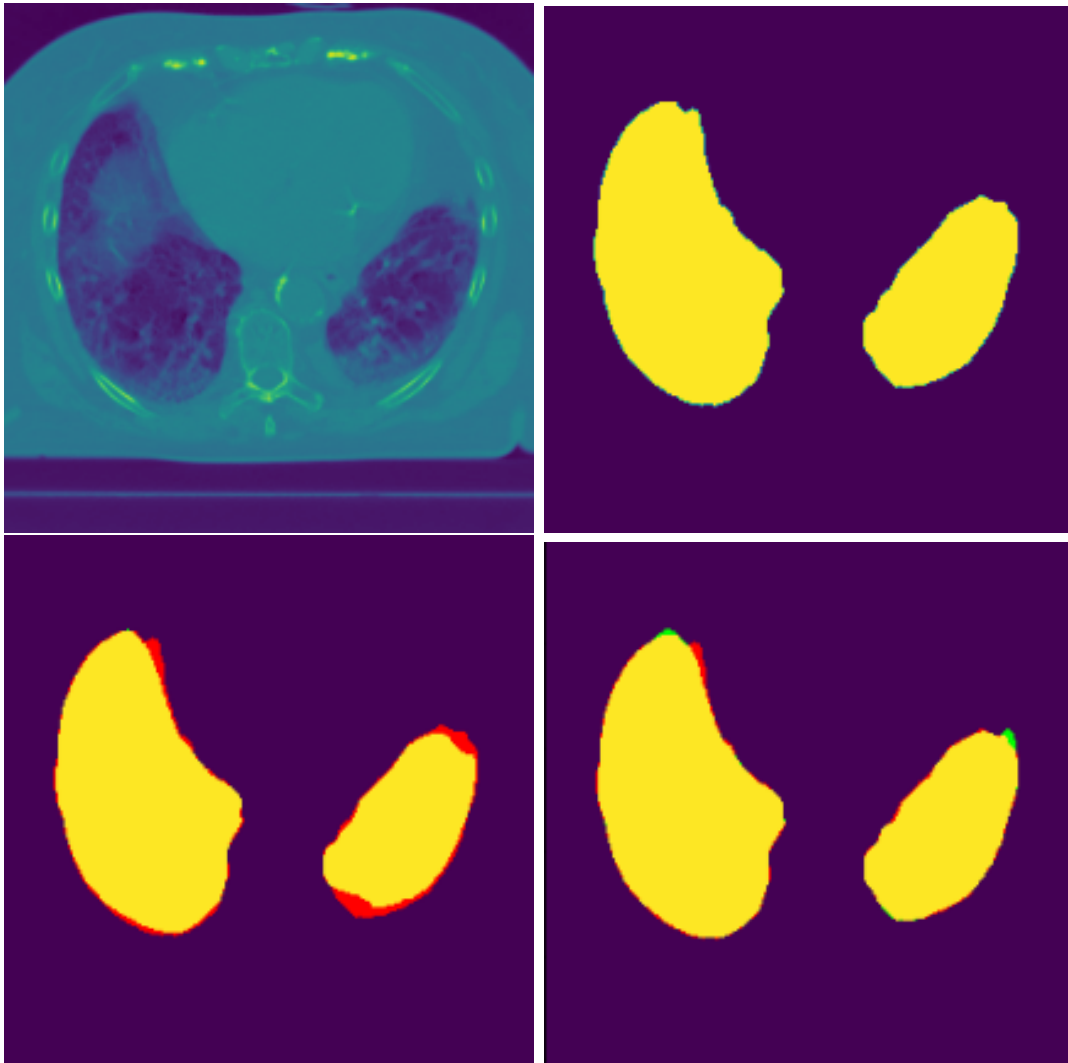
Orgaan	IoU score	Dicescore
Hart	0.86	0.92
Slokdarm	0.48	0.64
Longen	0.95	0.97
Ruggenmerg	0.60	0.70

Als we deze resultaten vergelijken met tabel 54 zien we dat onze resultaten afwisselend zijn. Voor het hart presteren we slechter dan de eerste twee onderzoeken en gelijkaardig aan het derde onderzoek. Voor de slokdarm presteren we beter dan het eerste onderzoek en slechter dan de andere twee. Voor de longen presteren we gelijkaardig. Voor de luchtpijp en het ruggenmerg presteren we slechter.

Indien we onze scores vergelijken met de variabiliteit tussen clinici via tabel 55 zien we echter dat we gelijkaardig presteren met waarden voor de IoU en Dicescores die zeer dicht bij de klinische variabiliteit ligt. Enkel het hart presteert relatief slecht. Een mogelijke verklaring is dat de resultaten van de onderzoeken afkomstig waren van sets met weinig variabiliteit. Dit was bijvoorbeeld al het geval voor onderzoek 1 waar er slechts één volledige CT-scan werd gebruikt. Deze was dan vermoedelijk geheel geannoteerd door dezelfde persoon.

### 3.5.1 Longen

Paars is TN, geel is TP, groen is FP en rood is FN. De eerste set is afkomstig van de trainpatiënt en de tweede set van de testpatiënt.

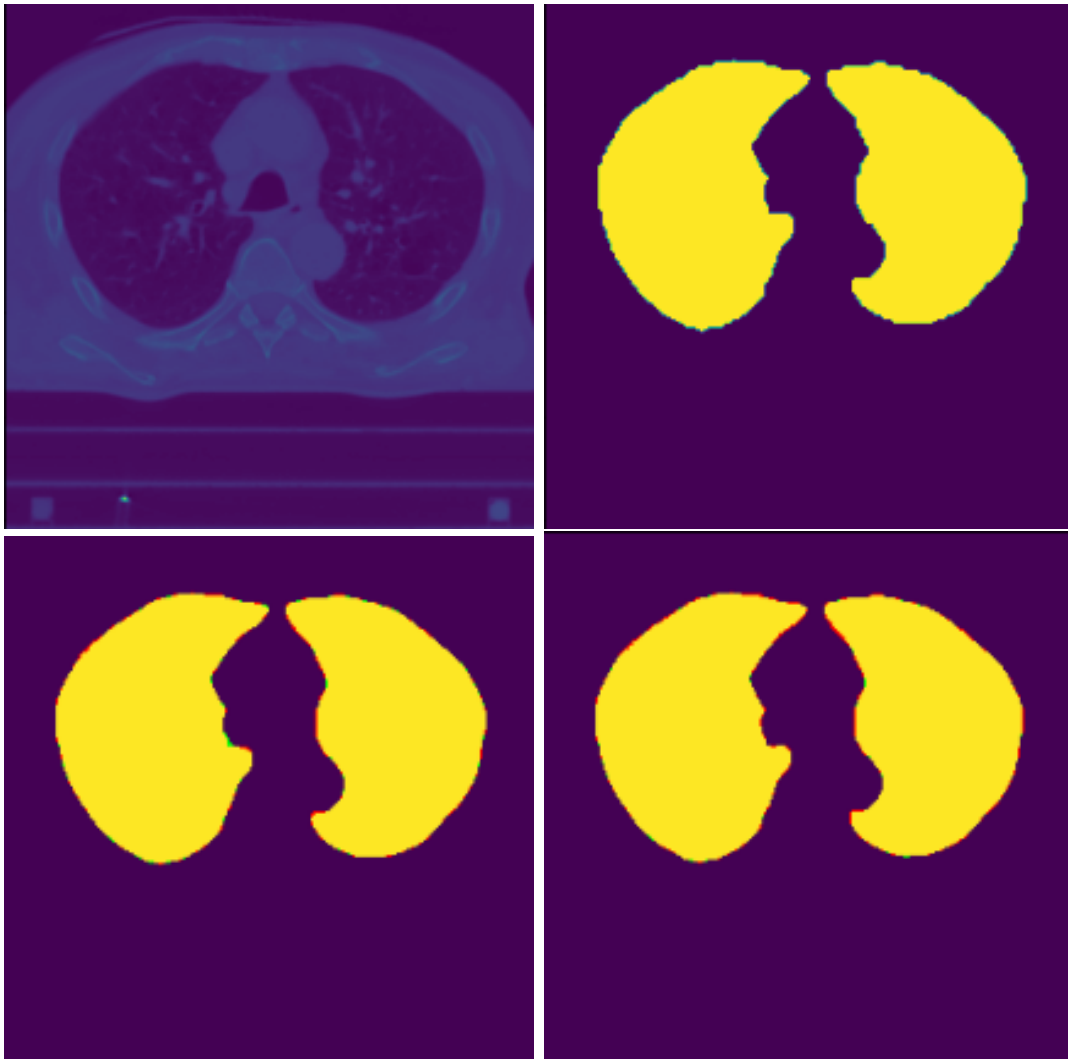


Figuur 81: CT-beeld (linksboven) ground truth long-segmentatie (rechtsboven) voorspelling configuratie 1 (linksonder) en voorspelling cofiguratie 2 (rechtsonder) van de trainpatiënt.

We zien een zeer sterke overeenkomst tussen het ware masker en de voorspellingen door beide configuraties. De grotere dataset heeft geholpen bij het beter intekenen van de longen, kijkende naar configuratie 1 waar er toch nog redelijk wat valse negatieven waren. Algemeen hebben we een zeer goede voorspelling. Dit is nodig aangezien het falen van het trainen van de longen zou betekenen dat het model zeker niet trainbaar is gezien de lage moeilijkheidsgraad van de longen.

We zien wel nog bepaalde randgebieden waar er discussie is tussen het model en de ground truth. We kunnen de discussie houden welke van de twee het meest correct is aangezien voor de longen vaak al een automatische intekening in Raystation gebeurt.





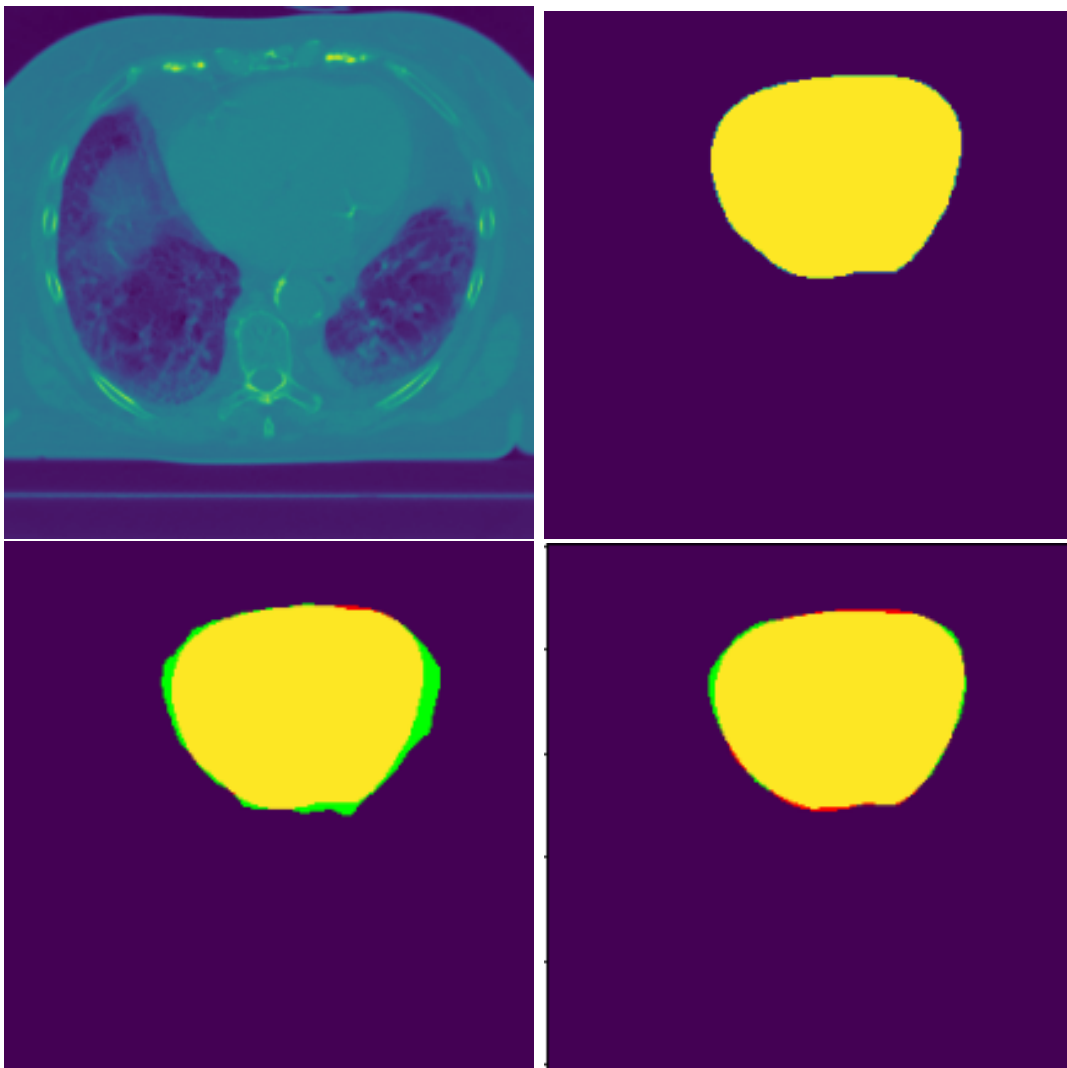
Figuur 82: CT-beeld (linksboven) ground truth long-segmentatie (rechtsboven) voorspelling configuratie 1 (linksonder) en voorspelling configuratie 2 (rechtsonder) van de testpatiënt.

Opnieuw zien we een bijna perfecte gelijkheid tussen het ware masker en de voorspellingen van beide configuraties.

Algemeen kunnen we concluderen dat de Dixeloss aanpak in staat is voldoende correcte voorspellingen te maken over de longen.

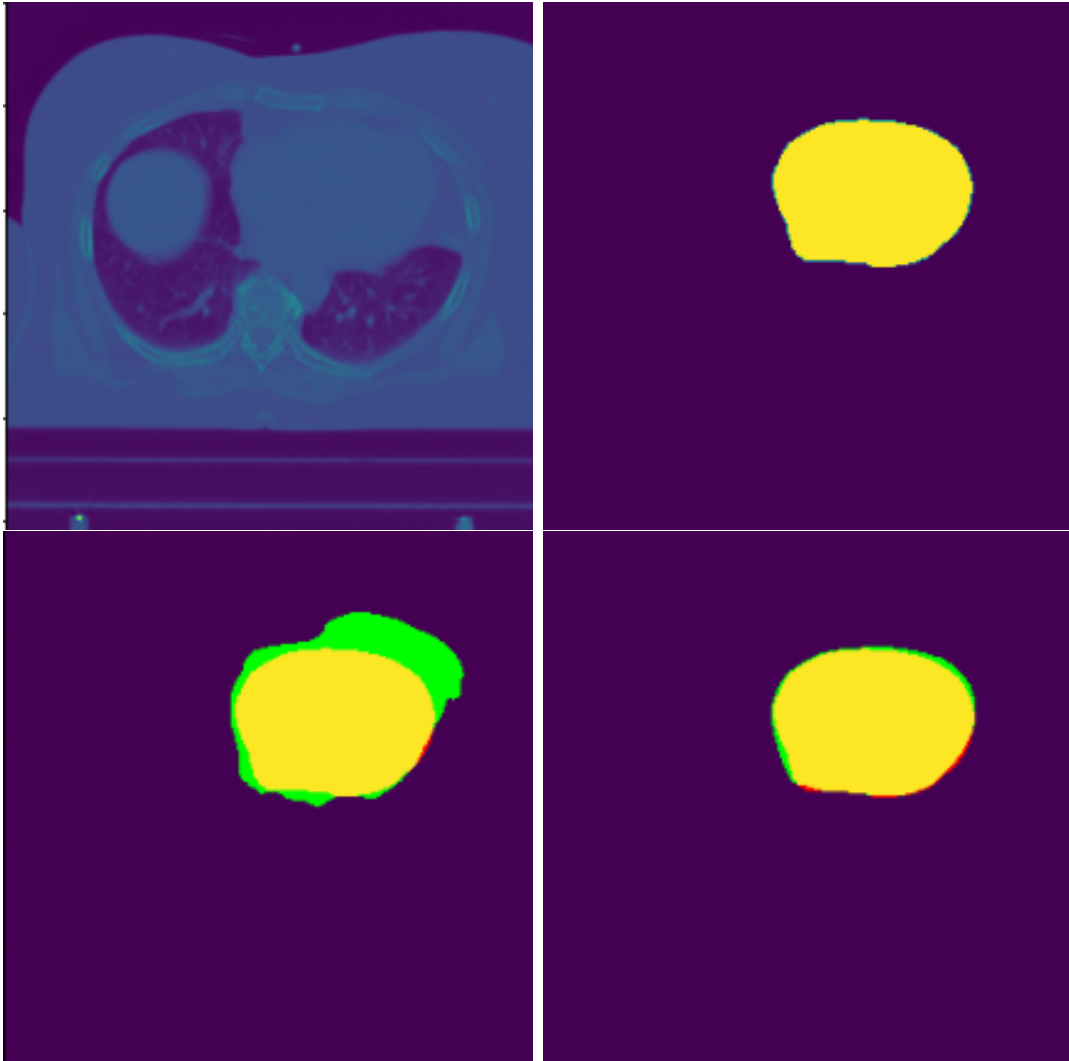
### 3.5.2 Hart

Paars is TN, geel is TP, groen is FP en rood is FN. De eerste set is afkomstig van de trainpatiënt en de tweede set van de testpatiënt.



Figuur 83: CT-beeld (linksboven) ground truth hart-segmentatie (rechtsboven) voorspelling configuratie 1 (linksonder) en voorspelling configuratie 2 (rechtsonder) van de trainpatiënt.

We zien bij configuratie 1 vooral valse positieven en, relatief gezien, kleine valse negatieven. De situatie is nagenoeg perfect bij configuratie 2. Dit wijst op het (over)trainbaar zijn van het hart.



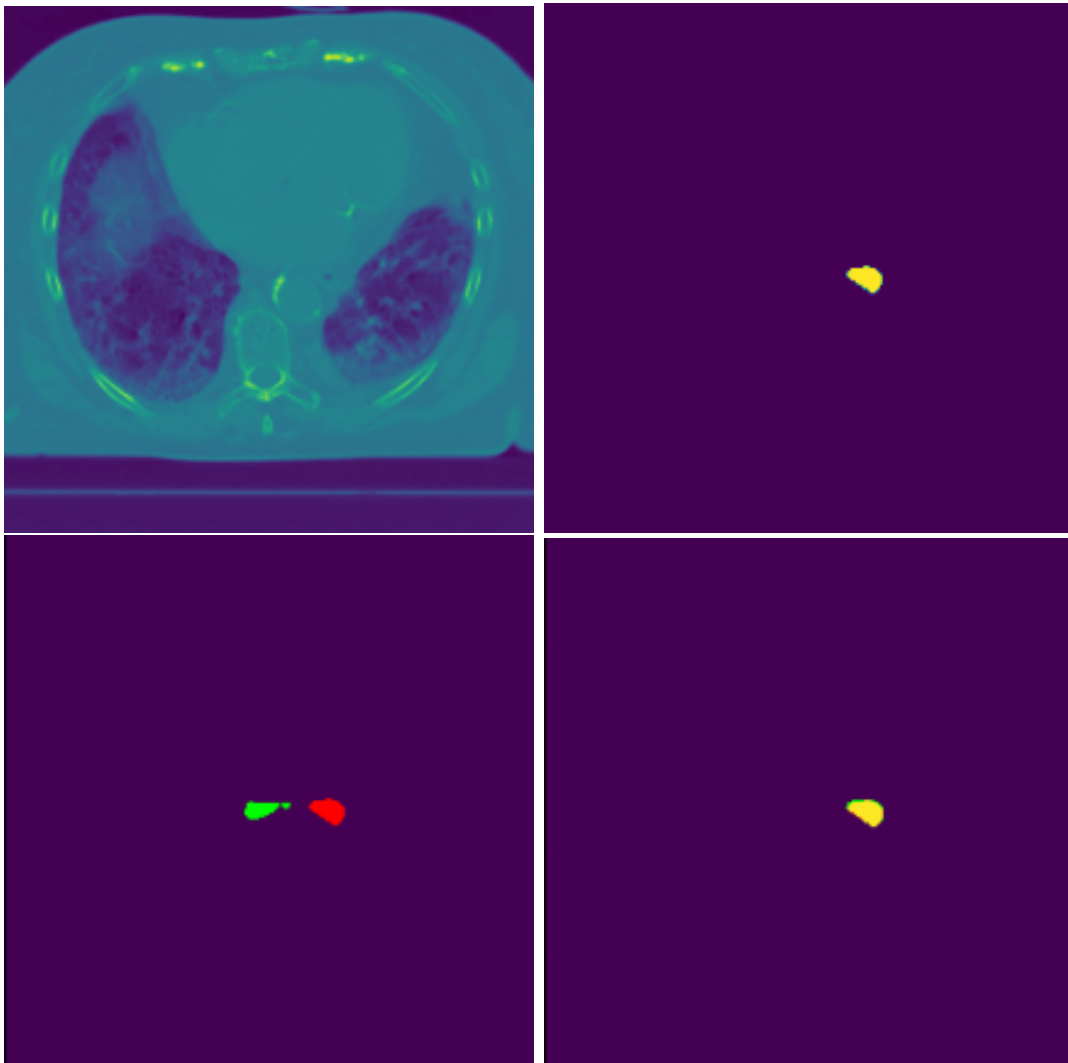
Figuur 84: CT-beeld (linksboven) ground truth hart-segmentatie (rechtsboven) voorspelling configuratie 1 (linksonder) en voorspelling configuratie 2 (rechtsonder) van de testpatiënt.

De valse positieven bij de testpatiënt bij configuratie 1 zijn wel uitzonderlijk groot. Alhoewel dit minder erg is dan valse positieven bij een orgaan is deze orde van fout wel extreem. De situatie bij configuratie 2 is echter uitstekend. Slechts minieme fouten waar het gross uit valse positieven bestaat.

We concluderen dat het Dice loss trainen mogelijk is voor het hart.

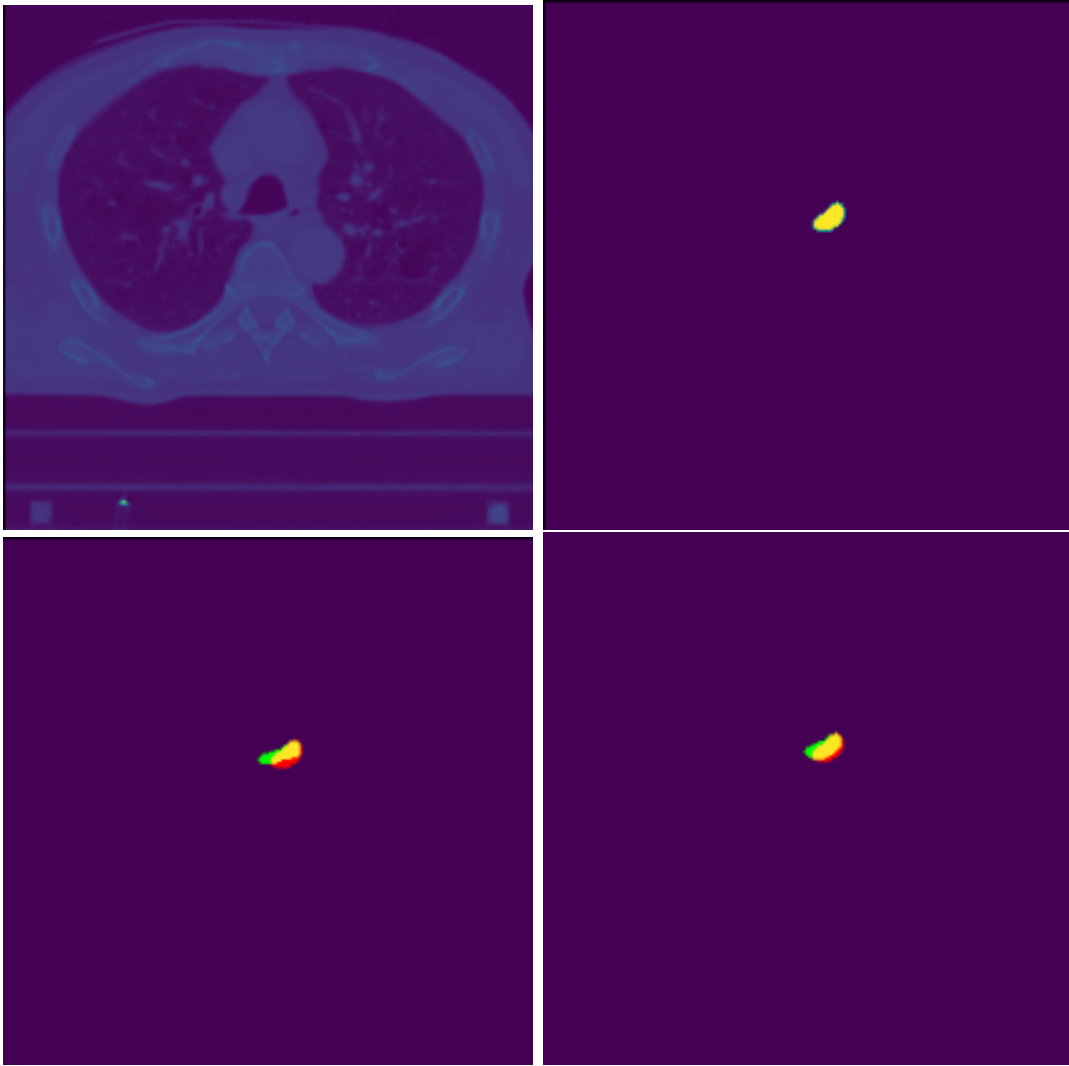
### 3.5.3 Slokdarm

Paars is TN, geel is TP, groen is FP en rood is FN. De eerste set is afkomstig van de trainpatiënt en de tweede set van de testpatiënt.



Figuur 85: CT-beeld (linksboven) ground truth slokdarm-segmentatie (rechtsboven) voorspelling configuratie 1 (linksonder) en voorspelling configuratie 2 (rechtsonder) van de trainpatiënt.

Bij configuratie 1 was het model volledig niet in staat het orgaan terug te vinden. Dit is het slechtst mogelijke scenario. De stralingsplanningsoftware zou op deze manier geen rekening houden met dit orgaan. Bij configuratie 2 zien we dat het model het orgaan relatief goed intekent. Er zijn slechts minieme fouten aan de rand die binnen de accuraatheid van de bestralingsapparatuur vallen.



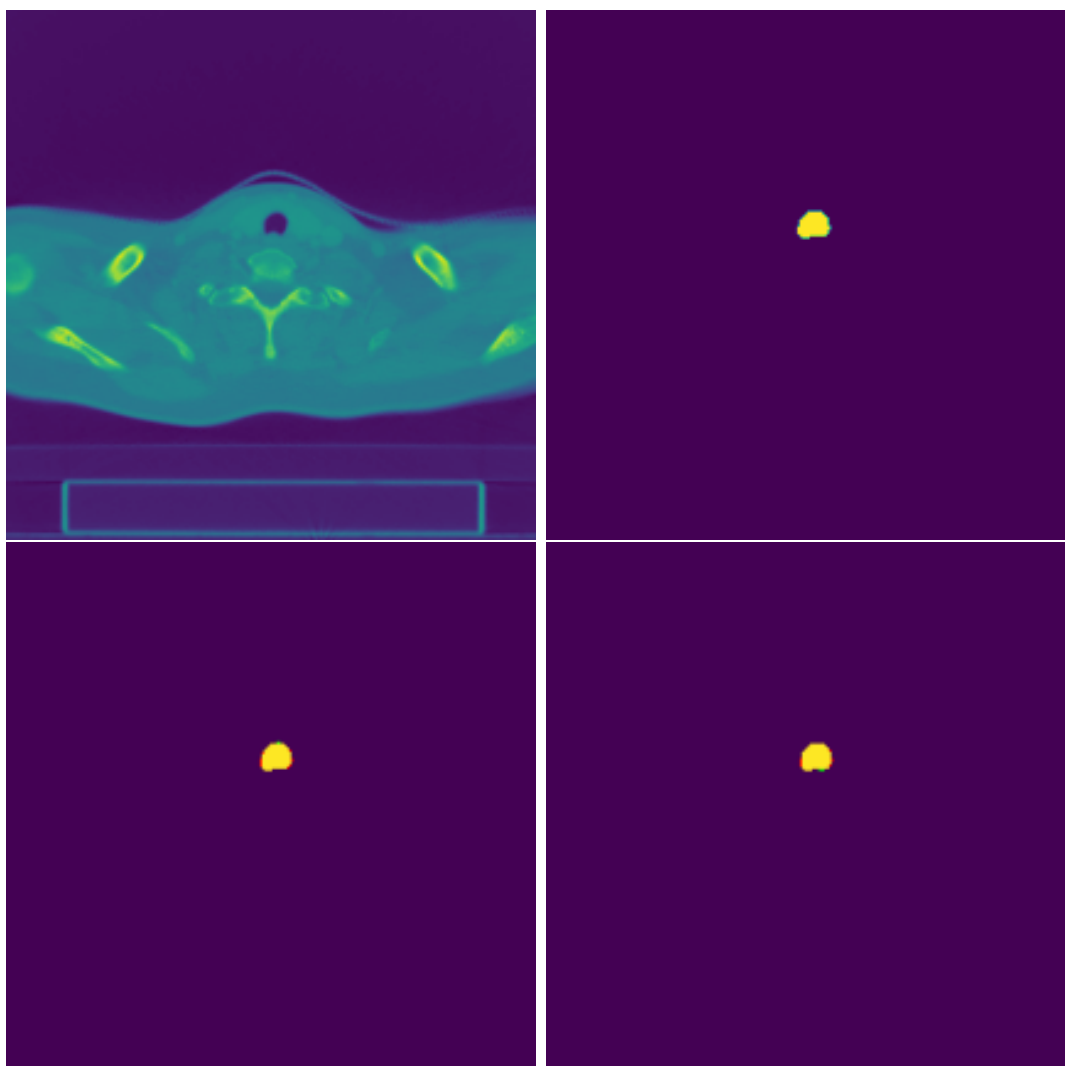
Figuur 86: CT-beeld (linksboven) ground truth slokdarm-segmentatie (rechtsboven) voorspelling configuratie 1 (linksonder) en voorspelling configuratie 2 (rechtsonder) van de testpatiënt.

De situatie van configuratie 2 bij de testpatiënt is iets minder ideaal maar komt nog steeds redelijk goed overeen. Als we de trend van configuratie 1 naar configuratie 2 zouden doortrekken, zouden we verwachten dat nog meer data de intekening verder zou verfijnen.

We concluderen dat het Dixeloss trainen op de slokdarm relatief wenselijk is, maar dat verdere verfijningen van het model door meer data het model ten goede zal komen.

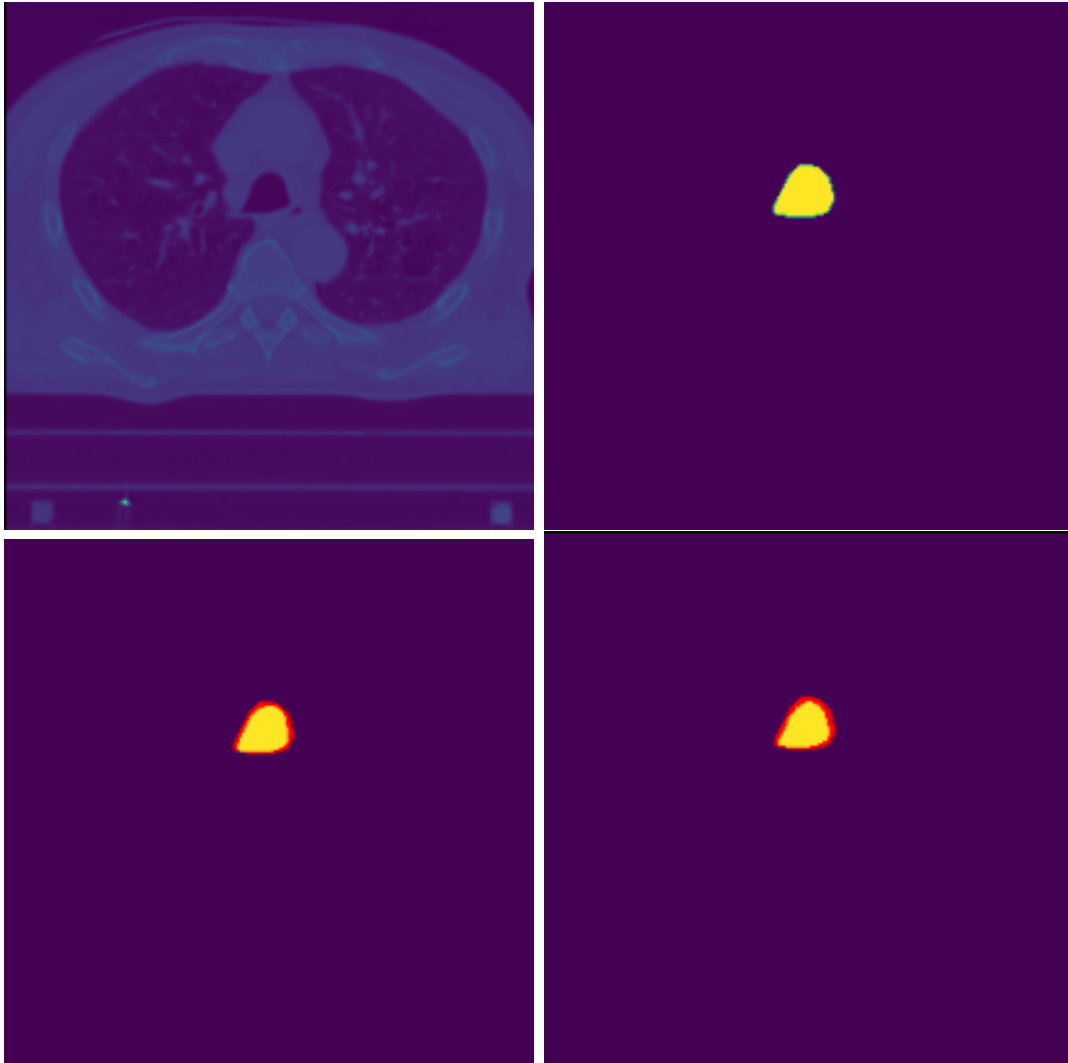
### 3.5.4 Luchtpijp

Paars is TN, geel is TP, groen is FP en rood is FN. De eerste set is afkomstig van de trainpatiënt en de tweede set van de testpatiënt.



Figuur 87: CT-beeld (linksboven) ground truth luchtpijp-segmentatie (rechtsboven) voorspelling configuratie 1 (linksonder) en voorspelling configuratie 2 (rechtsonder) van de trainpatiënt.

We zien dat het model zeer goed in staat is de luchtpijp terug te vinden. Ook zijn de fouten slechts van de orde van een pixel of twee. Het model is dus zeker trainbaar op dit orgaan. We kijken verder naar de testpatiënt om de generaliseerbaarheid te testen.



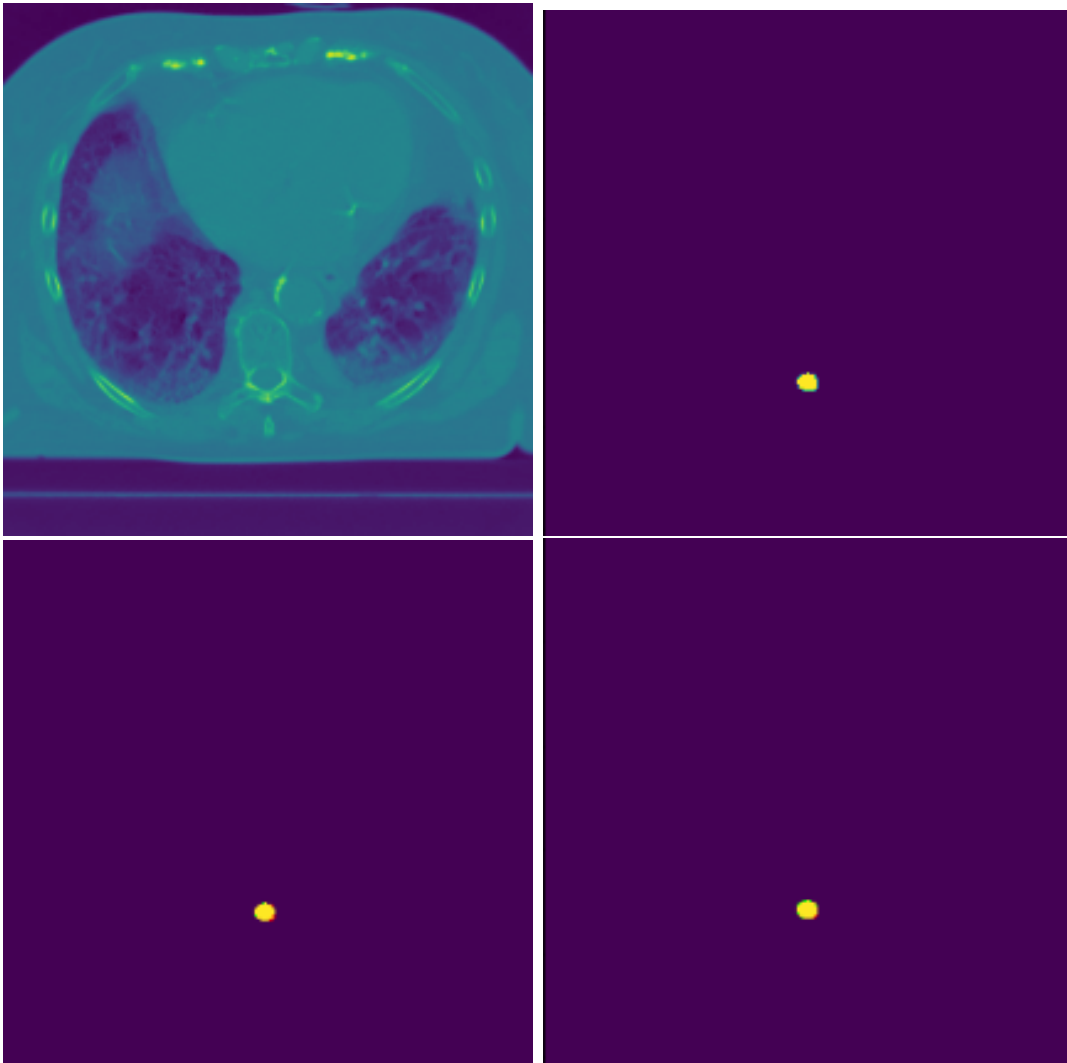
Figuur 88: CT-beeld (linksboven) ground truth luchtpijp-segmentatie (rechtsboven) voorspelling configuratie 1 (linksonder) en voorspelling configuratie 2 (rechtsonder) van de testpatiënt.

De testpatiënt presteert ongeveer hetzelfde voor configuratie 1 als voor configuratie 2. We hebben wel nog een relatief dikke valse negatieve rand wat niet wenselijk is voor een orgaan. Als we kijken naar de CT-foto zien we dat het ingetekend beeld ongeveer even groot is als de luchtcaviteit. Vermoedelijk wordt het orgaan altijd wat groter ingetekend zodat ook de rand van de luchtpijp zeker ingetekend is. Dit is iets dat het model moet kunnen leren.

We concluderen dat het Dixeloss trainen op de luchtpijp vrij wenselijk is, maar eventueel nog verfijnd kan worden.

### 3.5.5 Ruggenmerg

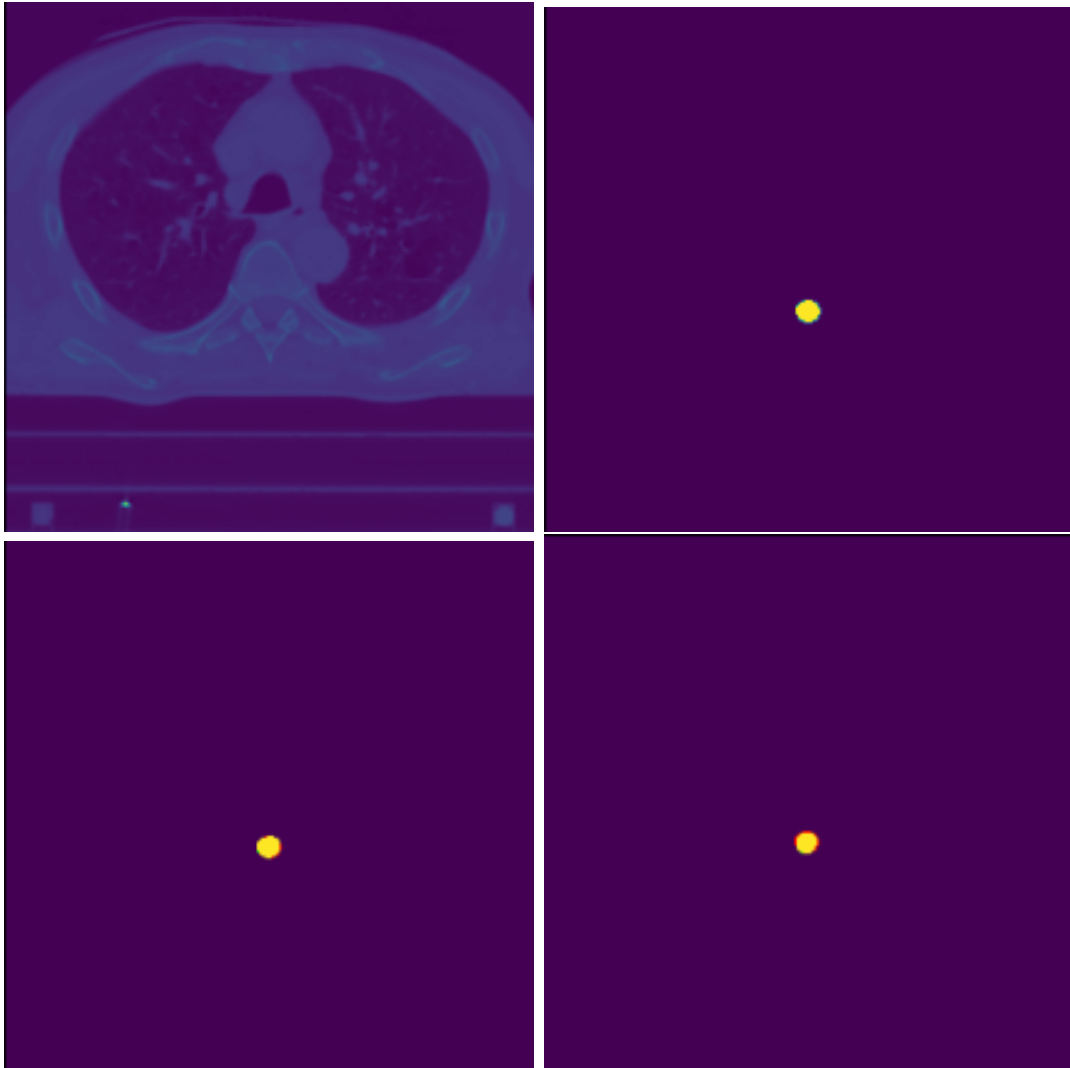
Paars is TN, geel is TP, groen is FP en rood is FN. De eerste set is afkomstig van de trainpatiënt en de tweede set van de testpatiënt.



Figuur 89: CT-beeld (linksboven) ground truth ruggenmerg-segmentatie (rechtsboven) voorspelling configuratie 1 (linksonder) en voorspelling configuratie 2 (rechtsonder) van de trainpatiënt.

We zien dat het model in beide situaties zeer goed in staat is het orgaan terug te vinden. Er is niet veel verbetering tussen configuratie 1 en 2 wat wijst op het al vroeg trainbaar zijn.





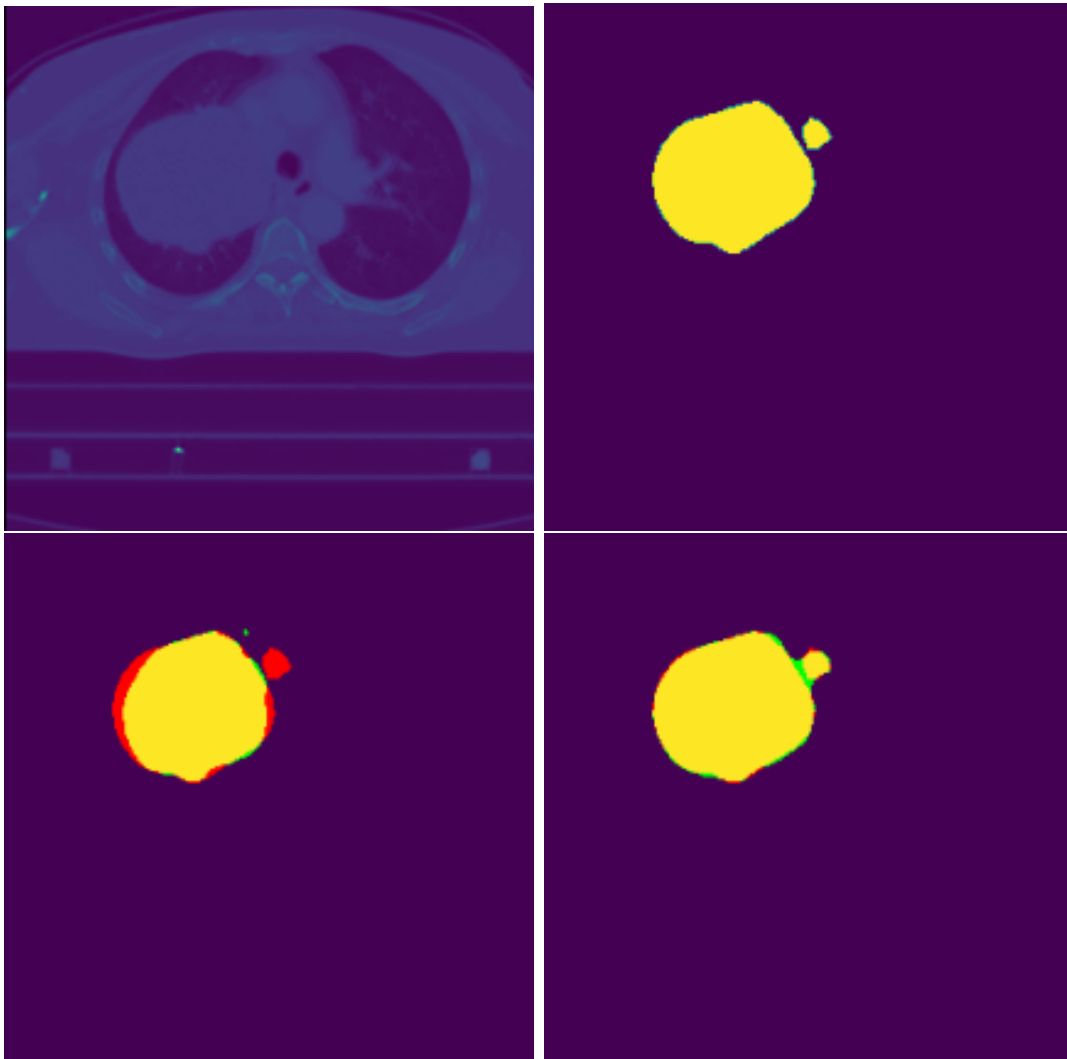
Figuur 90: CT-beeld (linksboven) ground truth ruggenmerg-segmentatie (rechtsboven) voorspelling configuratie 1 (linksonder) en voorspelling configuratie 2 (rechtsonder) van de testpatiënt.

Ook de testpatiënt presteert goed. Het model is ingetekend met slechts minieme randfouten die binnen de foutenmarge zal liggen.

We concluderen dat het Dixeloss trainen op het ruggenmerg zeer wenselijk is.

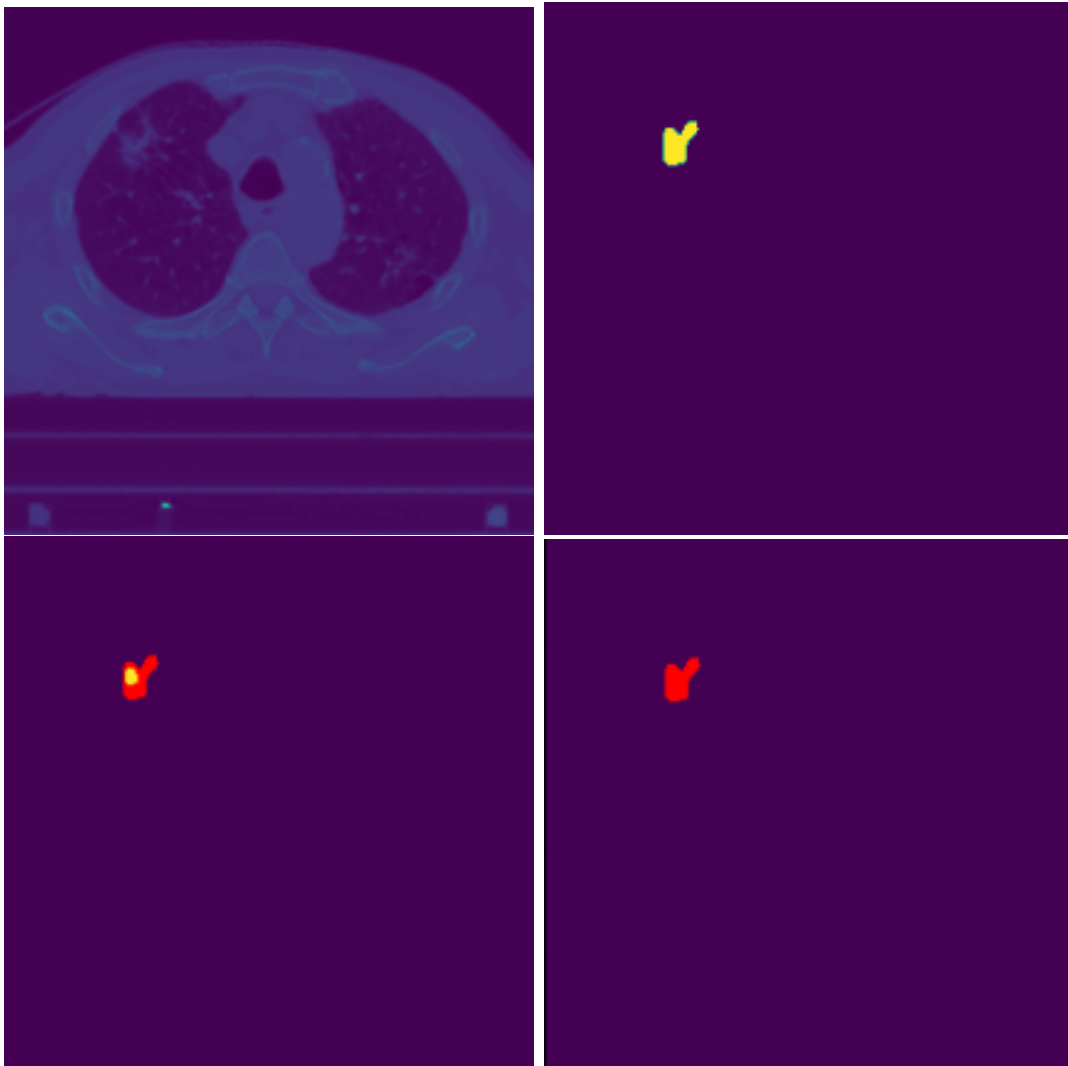
### 3.5.6 GTV

Paars is TN, geel is TP, groen is FP en rood is FN. De eerste set is afkomstig van de trainpatiënt en de tweede set van de testpatiënt.



Figuur 91: CT-beeld (linksboven) ground truth GTV-segmentatie (rechtsboven) voorspelling configuratie 1 (linksonder) en voorspelling configuratie 2 (rechtsonder) van de trainpatiënt.

We zien een verbetering van configuratie 1 naar configuratie 2. Echter zal deze foto even vaak zijn gepasseerd in beide configuraties. Dit wijst op een alsnog zekere gelijkens en zo ook trainbaar iets tussen verschillende tumoren aangezien de kennis voor een betere intekening alleen afkomstig kan geweest zijn van de grotere hoeveelheid aan tumorfoto's dat configuratie 2 heeft gezien.



Figuur 92: CT-beeld (linksboven) ground truth GTV-segmentatie (rechtsboven) voorspelling configuratie 1 (linksonder) en voorspelling configuratie 2 (rechtsonder) van de testpatiënt.

Er is een verslechtering van configuratie 1 naar 2. Dit wijst op het feit dat het model beter is geworden in het veralgemeniseren (wat ten goede komt voor de organen) maar zo ook slechter is geworden in het aanduiden van outliers.

We concluderen dat het Dixeloss trainen voor tumoren momenteel weinig potentie heeft. Indien we verbeteringen willen zien, zullen we veel meer data nodig hebben en het model beter moeten aanleren hoe gezond weefsel eruit ziet zodat het outliers kan detecteren.

### 3.6 DHV en clinical goals

Het model van configuratie 2 uit sectie 3.5 werd gebruikt om voorspellingen te doen op de vijf patiënten die samen de vijf testschijnpatiënten uitmaakten. Hierbij werd enkel het GTV niet meegenomen in de voorspelling wegens het compleet niet bruikbaar zijn. We herhalen dat enkel de centrale 256x256 pixels werden gebruikt van elke 512x512 foto. Helaas had één van de testpatiënten enkele stukken orgaan die buiten dit centrale beeld vielen. Daarom zullen enkel de overige vier testpatiënten gebruikt worden voor de DHV's en clinical goals. Voor elke patiënt zullen we twee stralingsplannen bekijken. Het originele dat gebaseerd is op de intekening van de radiotherapeuten-oncologen en hetgene bekomen door onze vijf ingetekende organen. Een beknopte bespreking bevindt zich in sectie 4.6.

Voor de stralingsplannen werden voor de originele en de testpatiënten dezelfde type afgeleide structuren gemaakt:

- Lungs-GTV. Dit zijn de longen zonder de pixels die binnen het GTV vallen. In ons geval zal het GTV in beide situaties hetgene zijn van de radiotherapeuten-oncologen.
- SpinalCord\_PRV05. Dit is een zogenaamde "Planning Risk Volume" (PRV) verkregen door een uniforme expansie van het ruggenmerg met 5 mm. Dit wordt gedaan bij de seriële risico-organen om te vermijden dat een kleine onzekerheid in de positionering van de patiënt tijdens de behandeling voor ongewenste effecten zou zorgen. Er wordt met andere woorden een zekere marge ingebouwd. De clinical goals moeten behaald worden voor het PRV.
- Esophagus\_PRV03. Hetzelfde principe als SpinalCord\_PRV05 maar dan met een marge van 3 mm.

### 3.6.1 Visuele vergelijking

Voor elk van de vier patiënten zal er een visuele vergelijking gegeven worden tussen de radiotherapeuten-oncologen hun intekeningen en onze eigen gegenereerde intekeningen. Voor alle testorganen werd ook een *expert opinion* opgemaakt door een dosimetrist van het Universitair Ziekenhuis Gent. Diens globale opinie was dat de resultaten zeker niet slecht waren en de afwijkingen voor de verschillende organen redelijk systematisch waren tussen de verschillende patiënten. De specifieke opmerkingen per patiënt en orgaan zullen nu besproken worden. In appendix A kan men de foto's aangehaald in deze sectie terugvinden.

#### 3.6.1.1 Patiënt 1

De dosimetrist had volgende geparafraseerde opmerkingen over de vijf organen.

Het hart was in orde met slechts grote afwijkingen bij het begin en het einde van het hart. Concreet was dit één slice bij het begin van het orgaan en vier slices bij het einde van het orgaan. In figuur 105 zien we vier foto's uit Raystation. Hierbij is de volle lijn de originele structuur en de stippellijn onze teststructuur.

De slokdarm was globaal in orde, maar niet in orde op bepaalde posities. Concreet was het niet in orde over drie series van CT-beelden bestaande uit negen, zes en 21 CT-beelden of dus in totaal 36 beelden van het totale aantal 99 CT-beelden waar er origineel een slokdarm was ingetekend. In figuur 106 zien we vier foto's uit Raystation. Hierbij is de volle lijn de originele structuur en de stippellijn de teststructuur.

De longen waren in orde met geen verdere opmerkingen. In figuur 107 zien we vier foto's uit Raystation. Hierbij is de volle lijn de originele structuur en de stippellijn de teststructuur.

Het ruggenmerg was in orde met de bijkomende opmerking dat het niet volledig was ingetekend. Concreet was het niet ingetekend over vier series van CT-beelden bestaande uit 13, 35, drie en drie CT-beelden of dus in totaal 54 beelden van het totale aantal 247 CT-beelden waar er origineel een ruggenmerg was ingetekend. Deze missende stukken bevonden zich echter ver in het begin en het einde van het ruggenmerg en daardoor ver van de hoge-dosisregio. In figuur 108 zien we vier foto's uit Raystation. Hierbij is de volle lijn de originele structuur en de stippellijn de teststructuur.

De luchtpijp was in orde met de bijkomende opmerking dat er twee CT-beelden bij het begin van de luchtpijp niet in orde waren. Dit zal echter niet het verschil maken. In figuur 109 zien we vier foto's uit Raystation. Hierbij is de volle lijn de originele structuur en de stippellijn de teststructuur.

### 3.6.1.2 Patiënt 2

De dosimetrist had volgende geparafraseerde opmerkingen over de vijf organen.

Het hart was in orde met slechts grote afwijkingen bij het begin en het einde van het hart. Concreet was dit twee slices bij het begin van het orgaan en vijf slices bij het einde van het orgaan. In figuur 110 zien we vier foto's uit Raystation. Hierbij is de volle lijn de originele structuur en de stippellijn de teststructuur.

De slokdarm was globaal in orde maar niet in orde op bepaalde posities. Concreet was het niet in orde over drie series van CT-beelden bestaande uit 18, 15 en zes slices of dus in totaal 39 beelden van het totale aantal 118 CT-beelden waar er origineel een slokdarm was ingetekend. Daarnaast voorspelden we acht extra stukken slokdarm bij het begin van het orgaan. In figuur 111 zien we vier foto's uit Raystation. Hierbij is de volle lijn de originele structuur en de stippellijn de teststructuur.

De longen waren in orde met slechts op bepaalde stukken een afwijking. Hierbij had het model stukken van de darmen ingetekend als long. Concreet was dit het geval in drie CT-beelden en waren de ingetekende regio's relatief klein. In figuur 112 zien we vier foto's uit Raystation. Hierbij is de volle lijn de originele structuur en de stippellijn de teststructuur.

Het ruggenmerg was in orde met de bijkomende opmerking dat het model het ruggenmerg verder heeft ingetekend dan de radiotherapeuten-oncologen. Concreet heeft het tien slices meer bij het begin van het orgaan en 38 slices meer bij het einde van het orgaan ingetekend. Het lijkt ons dat het ruggenmerg niet 100 procent volledig was ingetekend langs beide uiteinden. Wat in principe niet erg is omdat organen zo ver van de regio met de tumoren geen invloed zullen hebben op de stralingsplanning. In figuur 113 zien we vier foto's uit Raystation. Hierbij is de volle lijn de originele structuur en de stippellijn de teststructuur.

De luchtpijp was in orde met de bijkomende opmerking dat het model de luchtpijp verder heeft ingetekend dan de radiotherapeuten-oncologen. Concreet heeft het vijf slices meer bij het begin van het orgaan en twee slices meer bij het einde van het orgaan ingetekend. Bij het begin van het orgaan kunnen we het anders intekenen toeschrijven aan de anatomische overgang van de keel naar de luchtpijp. Bij het einde van het orgaan kunnen we het anders intekenen toeschrijven aan de anatomische overgang van de luchtpijp naar de twee bronchi. In figuur 114 zien we vier foto's uit Raystation. Hierbij is de volle lijn de originele structuur en de stippellijn de teststructuur.

### 3.6.1.3 Patiënt 3

De dosimetrist had volgende geparafraseerde opmerkingen over de vijf organen.

Het hart was in orde met slechts grote afwijkingen bij het begin en het einde van het hart. Concreet was dit voor één slice bij het begin van het orgaan en vier slices bij het einde van het orgaan. In figuur 115 zien we vier foto's uit Raystation. Hierbij is de volle lijn de originele structuur en de stippellijn de teststructuur.

De slokdarm was globaal in orde, maar niet in orde op bepaalde posities. Concreet waren er enkele slices (vier) waar het model, buiten de slokdarm, een extra stukje slokdarm voor-spelde. Daarnaast was er een serie van vijf CT-beelden waar de intekening niet compleet in orde was en een serie van 58 CT-beelden waar de intekening compleet niet in orde was. Dit geeft 63 van de 108 CT-beelden waar het intekenen niet in orde was. In figuur 116 zien we vier foto's uit Raystation. Hierbij is de volle lijn de originele structuur en de stippellijn de teststructuur.

De longen waren in orde met geen verdere opmerkingen. In figuur 117 zien we vier foto's uit Raystation. Hierbij is de volle lijn de originele structuur en de stippellijn de teststructuur.

Het ruggenmerg was in orde met de bijkomende opmerking dat het model bij het begin van het orgaan enkele keren te klein werd ingetekend. In figuur 118 zien we vier foto's uit Raystation. Hierbij is de volle lijn de originele structuur en de stippellijn de teststructuur.

De luchtpijp was in orde met de bijkomende opmerking dat het model driemaal een deel van de linker bronchus heeft ingetekend als luchtpijp. In figuur 119 zien we vier foto's uit Raystation. Hierbij is de volle lijn de originele structuur en de stippellijn de teststructuur.

#### 3.6.1.4 Patiënt 4

De dosimetrist had volgende geparafraseerde opmerkingen over de vijf organen.

Het hart was in orde met slechts grote afwijkingen bij het begin en het einde van het hart. Concreet was dit voor vijf slices bij het begin van het orgaan en drie slices bij het einde van het orgaan. In figuur 120 zien we vier foto's uit Raystation. Hierbij is de volle lijn de originele structuur en de stippellijn de teststructuur.

De slokdarm was globaal in orde maar niet in orde op bepaalde posities. Concreet waren er vier reeksen waar er afwijkingen waren van 18, vijf, vier en zeven slices. Dit geeft 34 van de 98 CT-beelden waar het intekenen niet in orde was. Dit bestond uit foto's waar het model een te kleine slokdarm voorspelde, de oppervlakten niet overlaptten of waar het model een tweede slokdarm voorspelde. In figuur 121 zien we vier foto's uit Raystation. Hierbij is de volle lijn de originele structuur en de stippellijn de teststructuur.

De longen waren in orde met geen verdere opmerkingen. Opvallend is dat het model soms nog kleine regio's van de longen wel intekende daar waar dit niet was gebeurd in het originele model. In andere patiënten zagen we deze kleine regio's soms wel origineel ingetekend. In figuur 122 zien we vier foto's uit Raystation. Hierbij is de volle lijn de originele structuur en de stippellijn de teststructuur.

Het ruggenmerg was in orde met de bijkomende opmerking dat er vier series van drie CT-beelden waren waar het model het ruggenmerg niet had ingetekend. Concreet waren dit drie regio's naar het einde van het ruggenmerg toe en één regio in het begin van het ruggenmerg. In figuur 123 zien we vier foto's uit Raystation. Hierbij is de volle lijn de originele structuur en de stippellijn de teststructuur.

De luchtpijp was in orde met de bijkomende opmerking dat het model lichtjes afweek in het begin en het einde van het orgaan. Bij het begin werden er twee slices niet ingetekend en bij het einde werden er twee slices meer ingetekend. Dit kadert zich weer in de discussie waar de overgang van keel naar luchtpijp is en waar de overgang van luchtpijp naar bronchi is. In figuur 124 zien we vier foto's uit Raystation. Hierbij is de volle lijn de originele structuur en de stippellijn de teststructuur.



### 3.6.2 Vergelijking van ROI (Region Of Interest) volumes.

Indien de organen relatief goed zijn ingetekend, zouden de opgemaakte volumes tussen de originele en de teststructuren relatief goed moeten overeenkomen. Kleine afwijkingen over een paar foto's zouden geen groot verschil mogen maken. Daarom zullen we de volumes voor elke testpatiënt en elk orgaan vergelijken met de originele volumes.

#### 3.6.2.1 Patiënt 1

Tabel 56: Vergelijking van ROI volumes van testpatiënt 1 tussen de originele en de teststructuren.

Orgaan	Origineel [cm <sup>3</sup> ]	Test [cm <sup>3</sup> ]	Absoluut verschil [cm <sup>3</sup> ]	Procentueel verschil [%]
Hart	627.63	618.31	9.32	1.48
Slokdarm	21.31	22.46	1.15	5.40
Longen	2054.89	2245.23	190.34	9.26
Ruggenmerg	42.71	42.47	0.24	0.56
Luchtpijp	17.26	21.92	4.66	27.00

We zien dat het verschil bij het hart en het ruggenmerg zeer klein is.

Het verschil bij de slokdarm valt nog mee in vergelijking met de visuele inspectie omdat de CT-slices met elkaar verbonden worden met als gevolg dat het ontbreken van enkele slices geen groot effect heeft. Het verschil blijft echter nog omdat het orgaan algemeen middelmatige afwijkingen kon hebben.

Het verschil bij de longen lijkt hoog. Dit is echter omdat de longen zelf al automatisch worden ingetekend met een threshold zoekend programma. Als we de foto's bekijken zou men kunnen oordelen dat ons programma het beter intekent. Dit is niet verrassend omdat ons programma veel complexer is dan het eenvoudige programma gebruikt in Raystation. Het feit dat de radiotherapeuten-oncologen deze afwijkingen van de automatisch ingetekende longen al tolereren wil zeggen dat onze intekeningen ook zeker in orde zijn, ondanks de procentuele afwijking van tien.

De luchtpijp heeft een relatief hoge afwijking. Dit komt omdat ons programma de luchtpijp consistent iets groter intekent. Aangezien de luchtpijp een relatief klein orgaan is (in doorsnede), kan een afwijking aan de rand relatief groot uitvallen.

### 3.6.2.2 Patiënt 2

Tabel 57: Vergelijking van ROI volumes van testpatiënt 2 tussen de originele en de teststructuren

Orgaan	Origineel [cm <sup>3</sup> ]	Test [cm <sup>3</sup> ]	Absoluut verschil [cm <sup>3</sup> ]	Procentueel verschil
Hart	837.20	797.49	39.71	4.74
Slokdarm	37.94	37.46	0.48	1.27
Longen	4903.58	4817.54	86.04	1.75
Ruggenmerg	64.14	53.8	10.34	16.12
Luchtpijp	68.68	54.43	14.25	20.75

We zien dat het verschil bij de slokdarm en de longen zeer klein is.

Het verschil bij het hart is te verklaren door het niet in orde zijn van de segmentaties die zich aan het begin en het einde van het hart bevonden. In realiteit valt dit dus redelijk mee.

Het verschil bij het ruggenmerg en de luchtpijp is relatief hoog. Dit komt omdat ons programma beide organen consistent iets kleiner intekent. Aangezien beide organen relatief kleine organen zijn (in doorsnede), kan een afwijking aan de rand relatief groot uitvallen.

### 3.6.2.3 Patiënt 3

Tabel 58: Vergelijking van ROI volumes van testpatiënt 3 tussen de originele en de teststructuren

Orgaan	Origineel [cm <sup>3</sup> ]	Test [cm <sup>3</sup> ]	Absoluut verschil [cm <sup>3</sup> ]	Procentueel verschil
Hart	893.63	688.59	205.04	22.94
Slokdarm	39.05	18.55	0.48	49.94
Longen	3326.74	3273.81	52.09	0.16
Ruggenmerg	70.15	58.93	11.22	15.99
Luchtpijp	53.00	36.48	16.52	31.17

We zien dat het verschil bij de longen zeer klein is.

Het verschil bij het hart is te verklaren door de grote verschillen bij de enkele slices in het begin en het einde van het orgaan.

Het verschil bij het ruggenmerg en de luchtpijp is relatief hoog. Dit komt omdat ons programma beide organen consistent iets kleiner intekent. Aangezien beide organen relatief kleine organen zijn (in doorsnede) kan een afwijking aan de rand relatief groot uitvallen.

Het verschil bij de slokdarm komt doordat het model bij een zeer groot aantal slices sterk afweek van de originele structuren.

### 3.6.2.4 Patiënt 4

Tabel 59: Vergelijking van ROI volumes van testpatiënt 4 tussen de originele en de teststructuren

Orgaan	Origineel [cm <sup>3</sup> ]	Test [cm <sup>3</sup> ]	Absoluut verschil [cm <sup>3</sup> ]	Procentueel verschil
Hart	668.38	637.89	28.49	4.26
Slokdarm	60.41	28.28	32.13	53.19
Longen	3824.05	3765.63	58.42	1.53
Ruggenmerg	57.63	56.20	1.43	2.48
Luchtpijp	39.74	42.6	2.86	7.20

We zien dat het verschil bij de longen en het ruggenmerg zeer klein is.

Het verschil bij het hart is te verklaren door de afwijkingen in de enkele slices bij het begin en het einde van het orgaan.

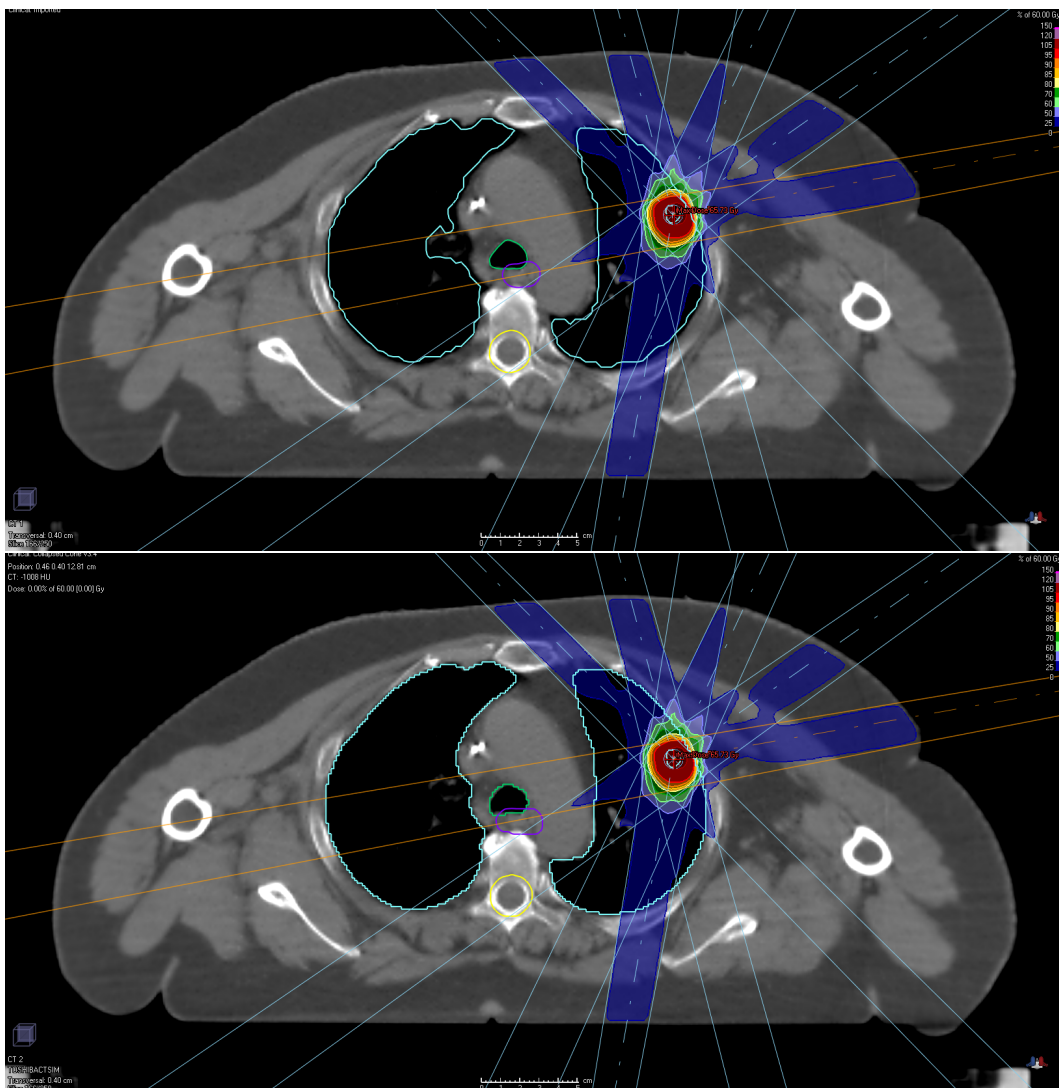
Het verschil bij de luchtpijp is te verklaren door het verschil in intekenen bij het einde van het orgaan dat een zeer grote doorsnede heeft.

Het extreem grote verschil bij de slokdarm is te verklaren door de grote afwijkingen bij het intekenen in verschillende secties van de slokdarm.

### 3.6.3 Dosisverdeling

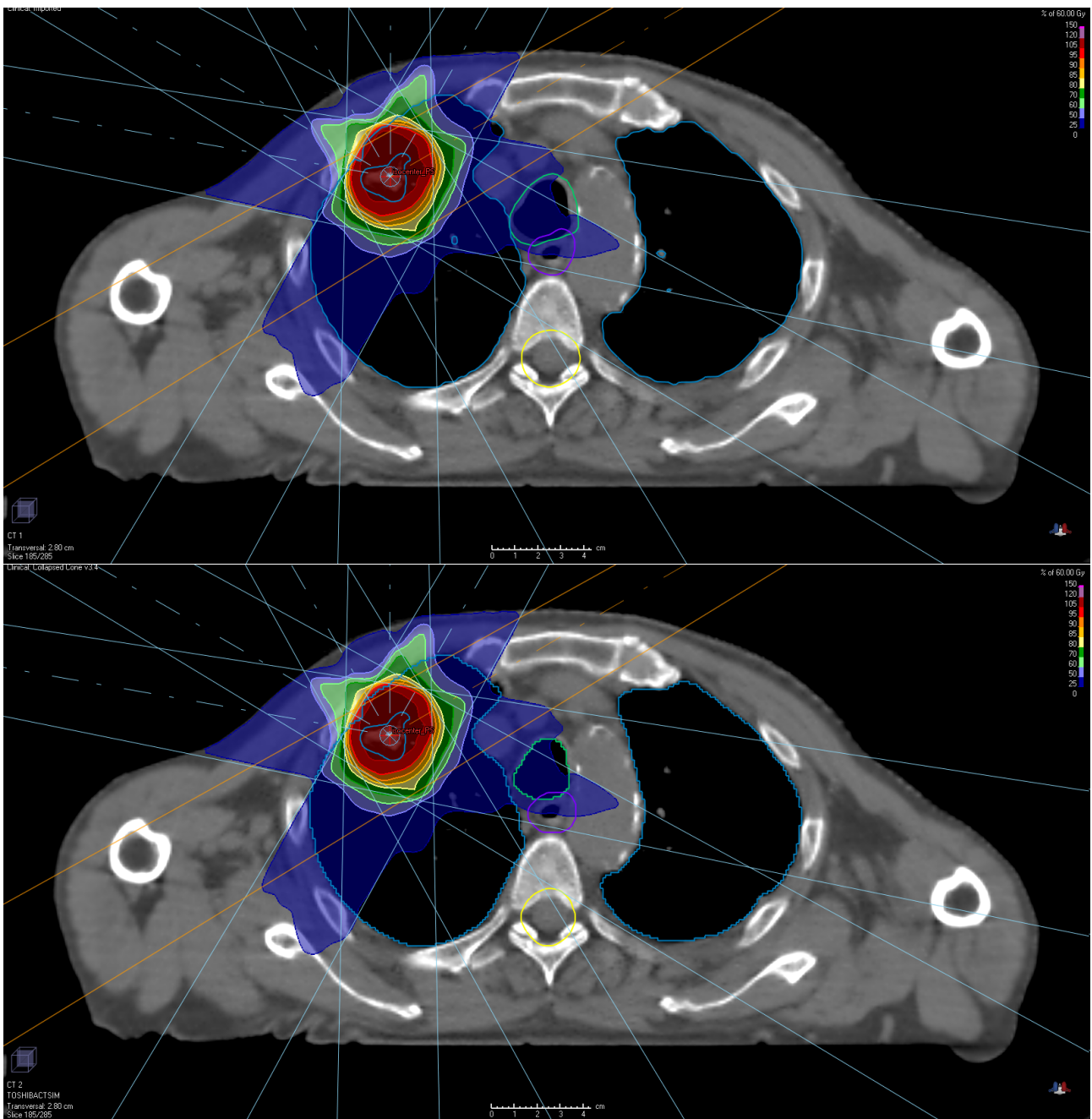
Nadat alle organen en tumoren zijn ingetekend, wordt er een stralingsplan opgesteld. Dit zoekt naar een zo hoog mogelijke dosis in de tumor en een zo laag mogelijke dosis in de organen. Raystation kan ons achteraf een visuele dosisverdeling leveren. Voor elke patiënt zullen we de dosisverdeling in de slice met het isocenter van de originele en de teststructuren tonen. Zo kan de lezer een beter beeld krijgen over de dosisverdeling. De getoonde dosisverdeling is degene berekend voor de stralingsplanning opgemaakt op basis van de originele structuren en is dus tweemaal identiek. Het verschil zal dus in de evaluatie zitten van deze dosisverdeling voor elke originele en teststructuur zoals de DVH's (sectie 3.6.4) en de clinical goals (sectie 3.6.5). Patienten één, twee en vier kregen een totale dosis van 60 Gy in drie fracties van 20 Gy. Patient drie kreeg een totale dosis van 66 Gy in 24 fracties van 2.75 Gy.

#### 3.6.3.1 Patiënt 1



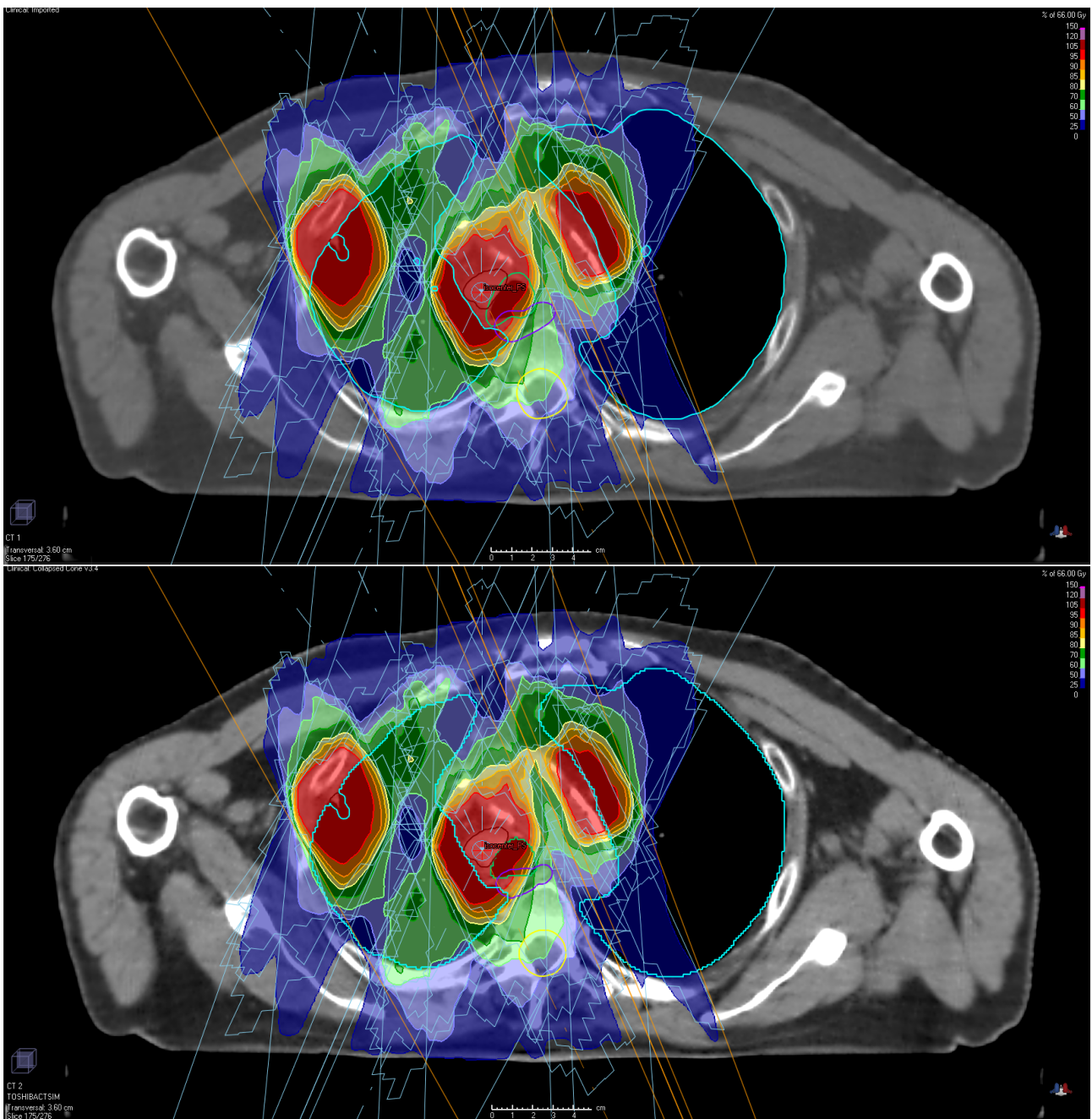
Figuur 93: Visuele dosisverdeling in de slice met het isocenter voor de originele (boven) en teststructuren (onder) van testpatiënt 1.

### 3.6.3.2 Patiënt 2



Figuur 94: Visuele dosisverdeling in de slice met het isocenter voor de originele (boven) en teststructuren (onder) van testpatiënt 2.

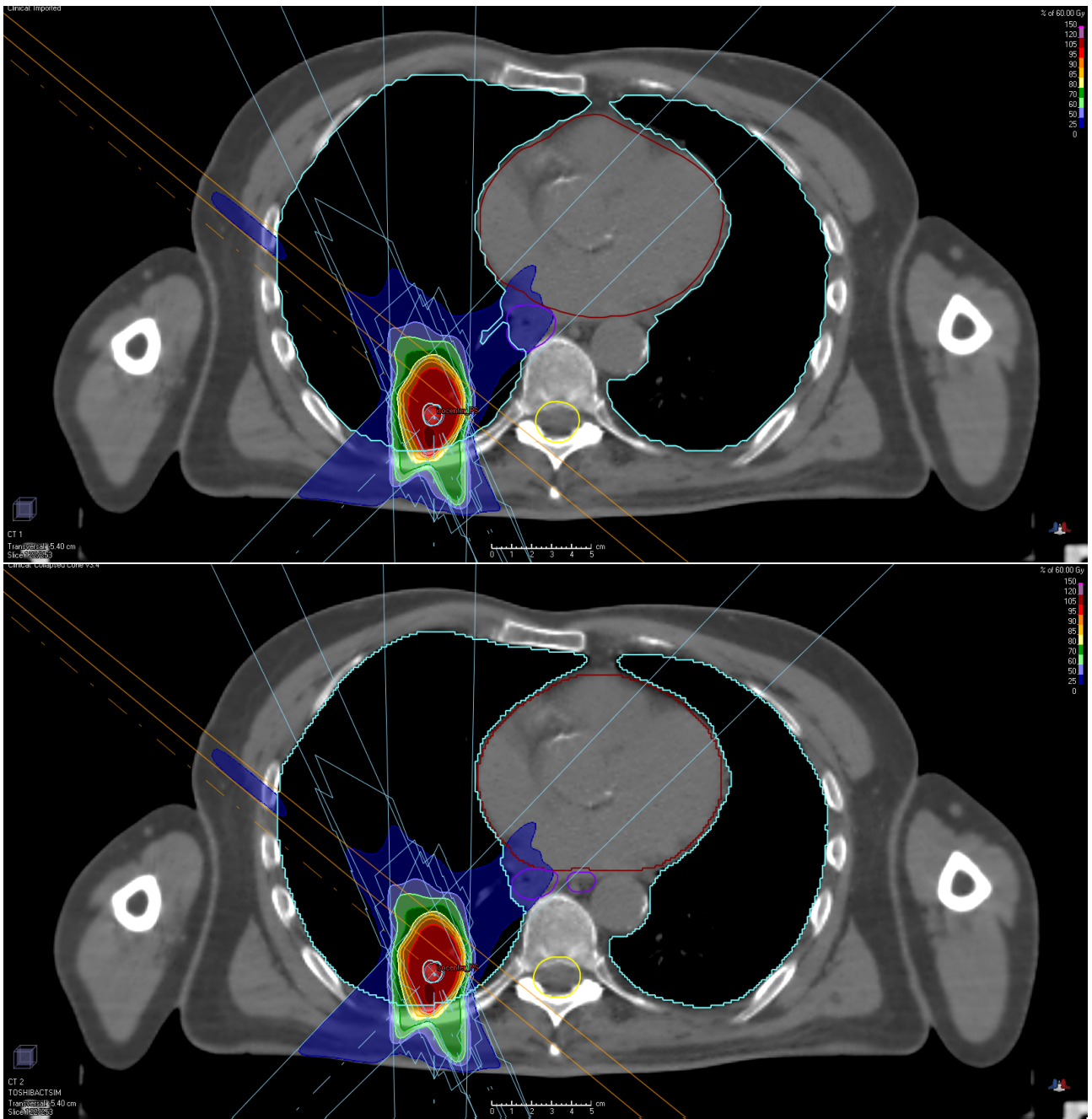
### 3.6.3.3 Patiënt 3



Figuur 95: Visuele dosisverdeling in de slice met het isocenter voor de originele (boven) en teststructuren (onder) van testpatiënt 3.



### 3.6.3.4 Patiënt 4

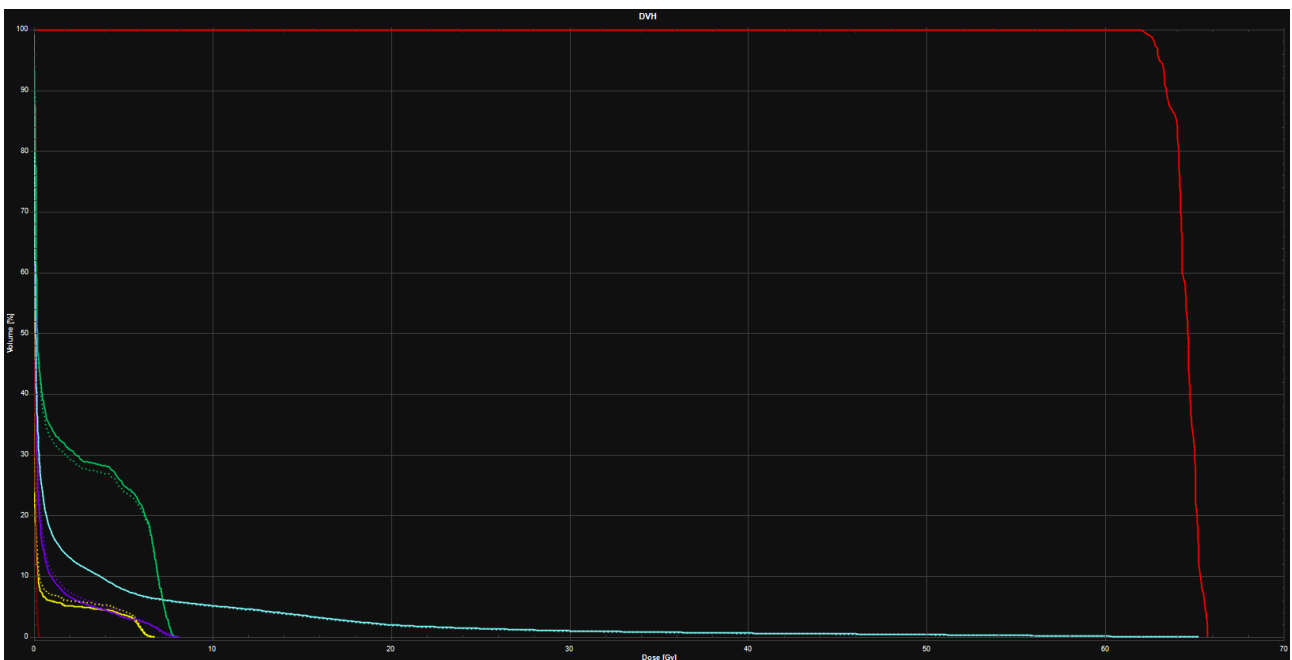


Figuur 96: Visuele dosisverdeling in de slice met het isocenter voor de originele (boven) en teststructuren (onder) van testpatiënt 4.

### 3.6.4 DVH

De DVH's (Dose Volume Histograms) vertellen ons hoeveel volumeprocent [%] een bepaalde dosis [Gy] of meer krijgt. Dit is opnieuw een maateenheid die klinisch relevant is. Indien deze histogrammen overeenkomen tussen de originele en de teststructuren kan men oordelen dat de intekeningen goed zijn gebeurd. We merken op dat we op zoek zijn naar histogrammen die goed overeen komen, niet histogrammen die eventueel beter of slechter zijn. Om dit te verduidelijken kijken we naar de hypothetische extreme situatie waar ons model de longen niet heeft ingetekend. Dan zouden we uiteraard geen dosis hebben in de longen. We willen dus toetsen hoe goed onze intekeningen overeen kwamen met die van de radiotherapeut-oncologen. De histogrammen zullen de vijf organen en het GTV bevatten. We laten het GTV staan ter illustratie van hoeveel dosis het GTV krijgt ten opzichte van de organen.

#### 3.6.4.1 Patiënt 1

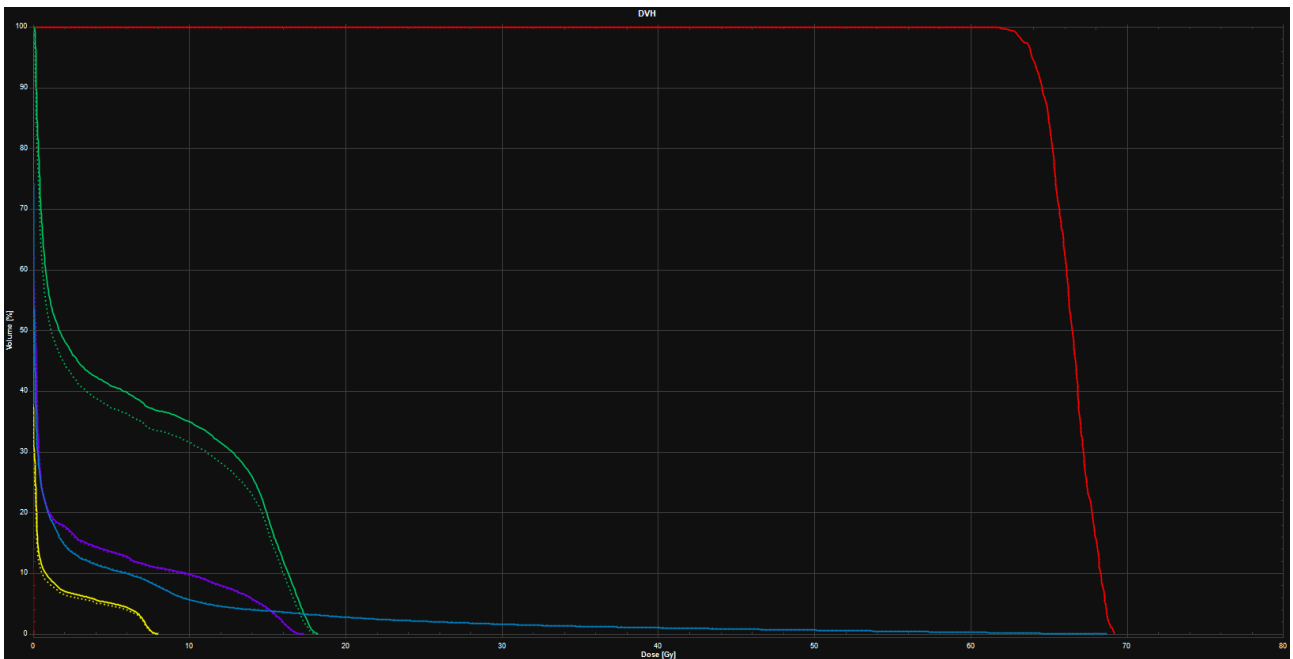


Figuur 97: DVH's voor testpatiënt één. Bordeaux: hart, rood: GTV (tumor), blauw: longen-GTV, paars: slokdarm\_PRV03, geelgroen: ruggenmerg\_PRV05, groen: luchtpijp. Volle lijn: originele structuren, stippellijn: teststructuren.

We zien dat de DVH's voor de longen, het hart en de slokdarm nagenoeg perfect overeenkomen. We merken een minimale afwijking op voor het ruggenmerg en een sterkere afwijking op voor de luchtpijp.



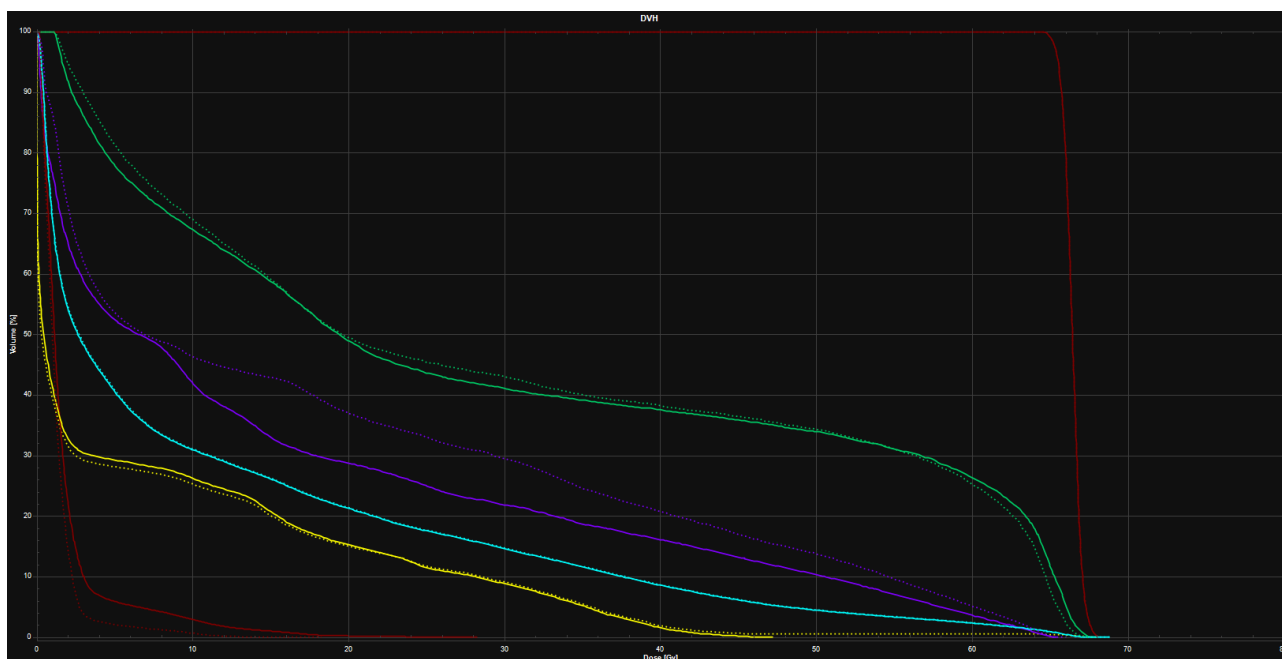
### 3.6.4.2 Patiënt 2



Figuur 98: DVH's voor testpatiënt twee. Bordeaux: hart, rood: GTV (tumor), blauw: longen-GTV, paars: slokdarm\_PRV03, geelgroen: ruggenmerg\_PRV05, groen: luchtpijp. Volle lijn: originele structuren, stippellijn: teststructuren.

We zien dat de DVH's voor de longen en het hart nagenoeg perfect overeenkomen. We merken een minimale afwijking op voor de slokdarm en slechts iets sterkere afwijkingen op voor de luchtpijp en het ruggenmerg.

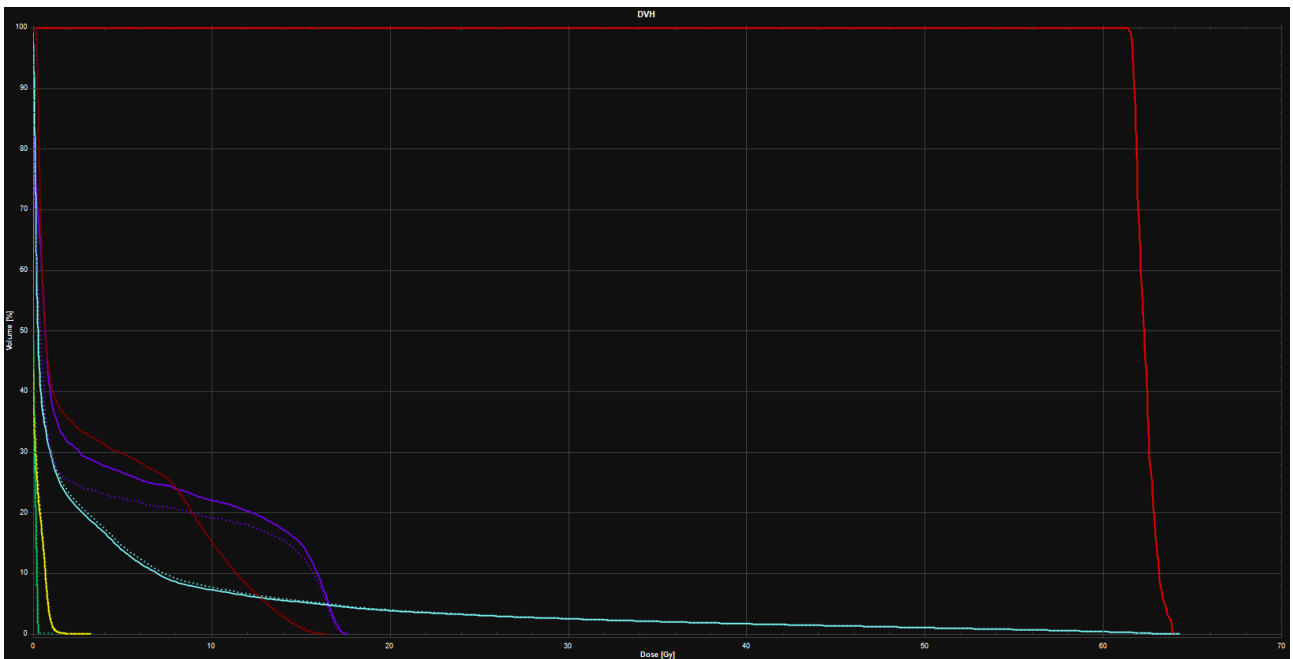
### 3.6.4.3 Patiënt 3



Figuur 99: DVH's voor testpatiënt drie. Bordeaux: hart, rood: GTV (tumor), blauw: longen-GTV, paars: slokdarm\_PRV03, geelgroen: ruggenmerg\_PRV05, groen: luchtpijp. Volle lijn: originele structuren, stippellijn: teststructuren.

We zien dat de DVH's voor de longen nagenoeg perfect overeenkomen. We merken een minimale afwijking op voor het ruggenmerg. Echter is deze minimale afwijking verraderlijk aangezien voor de testpatiënt de curve doorgaat tot in de orde van 60 Gy. Verder hebben we een sterkere afwijking voor het hart en de luchtpijp en een zeer sterke afwijking voor de slokdarm.

### 3.6.4.4 Patiënt 4



Figuur 100: DVH's voor testpatiënt drie. Bordeaux: hart, rood: GTV (tumor), blauw: longen-GTV, paars: slokdarm\_PRV03, geelgroen: ruggenmerg\_PRV05, groen: luchtpijp. Volle lijn: originele structuren, stippellijn: teststructuren.

We zien dat de DVH's voor alle organen buiten de slokdarm nagenoeg perfect overeenkomen. De slokdarm zelf heeft een redelijk grote afwijking.

### 3.6.5 Clinical goals

De clinical goals zijn zekere dosisrestricties opgelegd op de organen. Ze zeggen hoeveel volumeprocent van een orgaan maximaal een zekere dosis mag krijgen. Indien deze behaalde doelen overeenkomen tussen de originele en de teststructuren wil dit opnieuw zeggen dat er geen klinisch relevant verschil is tussen de twee intekeningen.









































#### 3.6.5.1 Patiënt 1

Priority ^	Dose	ROI/POI	Clinical goal	Value	Result
1	Plan dose: 2 Long SBR...	 Esophagus_PRV03	At most 24.00 Gy dose at 2.00 % volume	6.53 Gy	
1	Evaluation dose (2 Lon...	 Esophagus_PRV03	At most 24.00 Gy dose at 2.00 % volume	6.54 Gy	
1	Plan dose: 2 Long SBR...	 Heart	At most 24.00 Gy dose at 2.00 % volume	0.23 Gy	
1	Evaluation dose (2 Lon...	 Heart	At most 24.00 Gy dose at 2.00 % volume	0.22 Gy	
1	Plan dose: 2 Long SBR...	 Lungs-GTV	At most 20.00 Gy dose at 5.00 % volume	10.54 Gy	
1	Evaluation dose (2 Lon...	 Lungs-GTV	At most 20.00 Gy dose at 5.00 % volume	10.08 Gy	
1	Plan dose: 2 Long SBR...	 SpinalCord_PRV05	At most 18.00 Gy dose at 2.00 % volume	5.85 Gy	
1	Evaluation dose (2 Lon...	 SpinalCord_PRV05	At most 18.00 Gy dose at 2.00 % volume	5.91 Gy	
2	Plan dose: 2 Long SBR...	 Esophagus_PRV03	At most 27.00 Gy dose at 2.00 % volume	6.53 Gy	
2	Evaluation dose (2 Lon...	 Esophagus_PRV03	At most 27.00 Gy dose at 2.00 % volume	6.54 Gy	
2	Plan dose: 2 Long SBR...	 Heart	At most 26.00 Gy dose at 2.00 % volume	0.23 Gy	
2	Evaluation dose (2 Lon...	 Heart	At most 26.00 Gy dose at 2.00 % volume	0.22 Gy	
2	Plan dose: 2 Long SBR...	 Lungs-GTV	At most 20.00 Gy dose at 8.00 % volume	4.87 Gy	
2	Evaluation dose (2 Lon...	 Lungs-GTV	At most 20.00 Gy dose at 8.00 % volume	4.90 Gy	
2	Plan dose: 2 Long SBR...	 SpinalCord_PRV05	At most 22.00 Gy dose at 2.00 % volume	5.85 Gy	
2	Evaluation dose (2 Lon...	 SpinalCord_PRV05	At most 22.00 Gy dose at 2.00 % volume	5.91 Gy	

Figuur 101: Oplijsting van alle clinical goals (patiënt 1) en of ze al dan niet behaald zijn per Region of Interest. Plan dose is afkomstig van de originele structuren en evaluation dose is afkomstig van de teststructuren.

We merken op dat alle (afgeleide) teststructuren zijn geslaagd voor alle clinical goals. Daarbijkomend zijn de behaalde dosissen in de volumepercentages zeer gelijkend tussen de originele en de teststructuren.

### 3.6.5.2 Patiënt 2

Priority	Dose	ROI/POI	Clinical goal	Value	Result
1	Plan dose: 1_LONG_S...	 Esophagus_PRV03	At most 24.00 Gy dose at 2.00 % volume	15.95 Gy	
1	Evaluation dose (1_LO...	 Esophagus_PRV03	At most 24.00 Gy dose at 2.00 % volume	15.90 Gy	
1	Plan dose: 1_LONG_S...	 Heart	At most 24.00 Gy dose at 2.00 % volume	0.05 Gy	
1	Evaluation dose (1_LO...	 Heart	At most 24.00 Gy dose at 2.00 % volume	0.04 Gy	
1	Plan dose: 1_LONG_S...	 Lungs-GTV	At most 20.00 Gy dose at 5.00 % volume	11.03 Gy	
1	Evaluation dose (1_LO...	 Lungs-GTV	At most 20.00 Gy dose at 5.00 % volume	11.12 Gy	
1	Plan dose: 1_LONG_S...	 SpinalCord_PRV05	At most 18.00 Gy dose at 2.00 % volume	7.11 Gy	
1	Evaluation dose (1_LO...	 SpinalCord_PRV05	At most 18.00 Gy dose at 2.00 % volume	7.06 Gy	
1	Plan dose: 1_LONG_S...	 Trachea	At most 30.00 Gy dose at 2.00 % volume	17.57 Gy	
1	Evaluation dose (1_LO...	 Trachea	At most 30.00 Gy dose at 2.00 % volume	17.33 Gy	
2	Plan dose: 1_LONG_S...	 Esophagus_PRV03	At most 27.00 Gy dose at 2.00 % volume	15.95 Gy	
2	Evaluation dose (1_LO...	 Esophagus_PRV03	At most 27.00 Gy dose at 2.00 % volume	15.90 Gy	
2	Plan dose: 1_LONG_S...	 Heart	At most 26.00 Gy dose at 2.00 % volume	0.05 Gy	
2	Evaluation dose (1_LO...	 Heart	At most 26.00 Gy dose at 2.00 % volume	0.04 Gy	
2	Plan dose: 1_LONG_S...	 Lungs-GTV	At most 20.00 Gy dose at 8.00 % volume	7.90 Gy	
2	Evaluation dose (1_LO...	 Lungs-GTV	At most 20.00 Gy dose at 8.00 % volume	7.97 Gy	
2	Plan dose: 1_LONG_S...	 SpinalCord_PRV05	At most 22.00 Gy dose at 2.00 % volume	7.11 Gy	
2	Evaluation dose (1_LO...	 SpinalCord_PRV05	At most 22.00 Gy dose at 2.00 % volume	7.06 Gy	
2	Plan dose: 1_LONG_S...	 Trachea	At most 32.00 Gy dose at 2.00 % volume	17.57 Gy	
2	Evaluation dose (1_LO...	 Trachea	At most 32.00 Gy dose at 2.00 % volume	17.33 Gy	

Figuur 102: Oplijsting van alle clinical goals (patiënt 2) en of ze al dan niet behaald zijn per Region of Interest. Plan dose is afkomstig van de originele structuren en evaluation dose is afkomstig van de teststructuren.

We merken op dat alle (afgeleide) teststructuren zijn geslaagd voor alle clinical goals. Daarbijkomend zijn de behaalde dosissen in de volumepercentages zeer gelijkend tussen de originele en de teststructuren.

### 3.6.5.3 Patiënt 3





















Priority ^	Dose	ROI/POI	Clinical goal	Value	Result
1	Plan dose: 1 Long prim...	Lungs-GTV	At most 16.00 Gy average dose	11.19 Gy	
1	Evaluation dose (1 Lon...	Lungs-GTV	At most 16.00 Gy average dose	11.26 Gy	
1	Plan dose: 1 Long prim...	SpinalCord	At most 50.00 Gy dose at 0.00 % volume	45.77 Gy	
1	Evaluation dose (1 Lon...	SpinalCord	At most 50.00 Gy dose at 0.00 % volume	67.16 Gy	
2	Plan dose: 1 Long prim...	Lungs-GTV	At most 20.00 Gy average dose	11.19 Gy	
2	Evaluation dose (1 Lon...	Lungs-GTV	At most 20.00 Gy average dose	11.26 Gy	
2	Plan dose: 1 Long prim...	SpinalCord	At most 52.00 Gy dose at 0.00 % volume	45.77 Gy	
2	Evaluation dose (1 Lon...	SpinalCord	At most 52.00 Gy dose at 0.00 % volume	67.16 Gy	
	Plan dose: 1 Long prim...	Esophagus	At most 50.00 Gy dose at 50.00 % volume	6.02 Gy	
	Evaluation dose (1 Lon...	Esophagus	At most 50.00 Gy dose at 50.00 % volume	6.87 Gy	
	Plan dose: 1 Long prim...	Esophagus_PRV03	At most 66.00 Gy dose at 0.00 % volume	65.44 Gy	
	Evaluation dose (1 Lon...	Esophagus_PRV03	At most 66.00 Gy dose at 0.00 % volume	66.02 Gy	
	Plan dose: 1 Long prim...	GTV	At least 62.70 Gy dose at 95.00 % volume	65.53 Gy	
	Evaluation dose (1 Lon...	GTV	At least 62.70 Gy dose at 95.00 % volume	65.54 Gy	
	Plan dose: 1 Long prim...	GTV	At most 70.62 Gy dose at 2.00 % volume	67.58 Gy	
	Evaluation dose (1 Lon...	GTV	At most 70.62 Gy dose at 2.00 % volume	67.58 Gy	
	Plan dose: 1 Long prim...	Lungs-GTV	At most 20.00 Gy dose at 30.00 % volume	10.98 Gy	
	Evaluation dose (1 Lon...	Lungs-GTV	At most 20.00 Gy dose at 30.00 % volume	11.12 Gy	
	Plan dose: 1 Long prim...	Lungs-GTV	At most 30.00 Gy dose at 18.00 % volume	24.29 Gy	
	Evaluation dose (1 Lon...	Lungs-GTV	At most 30.00 Gy dose at 18.00 % volume	24.46 Gy	

Figuur 103: Oplijsting van alle clinical goals (patiënt 3) en of ze al dan niet behaald zijn per Region of Interest. Plan dose is afkomstig van de originele structuren en evaluation dose is afkomstig van de teststructuren.

We merken op dat de afgeleide teststructuur van de slokdarm net niet slaagt voor de clinical goal. Het kleine verschil in dosis is net genoeg om de grens te overschrijden.

De clinical goals voor het ruggenmerg falen drastisch. De dosis verschilt sterk met de originele dosis.

### 3.6.5.4 Patiënt 4

Priority	Dose	ROI/POI	Clinical goal	Value	Result
1	Plan dose: 2 LONG SB...	 Esophagus_PRV03	At most 24.00 Gy dose at 2.00 % volume	16.96 Gy	✓
1	Evaluation dose (2 LO...	 Esophagus_PRV03	At most 24.00 Gy dose at 2.00 % volume	16.84 Gy	✓
1	Plan dose: 2 LONG SB...	 Heart	At most 24.00 Gy dose at 2.00 % volume	14.42 Gy	✓
1	Evaluation dose (2 LO...	 Heart	At most 24.00 Gy dose at 2.00 % volume	14.66 Gy	✓
1	Plan dose: 2 LONG SB...	 Lungs-GTV	At most 20.00 Gy dose at 5.00 % volume	15.80 Gy	✓
1	Evaluation dose (2 LO...	 Lungs-GTV	At most 20.00 Gy dose at 5.00 % volume	16.37 Gy	✓
1	Plan dose: 2 LONG SB...	 SpinalCord_PRV05	At most 18.00 Gy dose at 2.00 % volume	1.01 Gy	✓
1	Evaluation dose (2 LO...	 SpinalCord_PRV05	At most 18.00 Gy dose at 2.00 % volume	1.04 Gy	✓
1	Plan dose: 2 LONG SB...	 Trachea	At most 30.00 Gy dose at 2.00 % volume	0.27 Gy	✓
1	Evaluation dose (2 LO...	 Trachea	At most 30.00 Gy dose at 2.00 % volume	0.30 Gy	✓
2	Plan dose: 2 LONG SB...	 Esophagus_PRV03	At most 27.00 Gy dose at 2.00 % volume	16.96 Gy	✓
2	Evaluation dose (2 LO...	 Esophagus_PRV03	At most 27.00 Gy dose at 2.00 % volume	16.84 Gy	✓
2	Plan dose: 2 LONG SB...	 Heart	At most 26.00 Gy dose at 2.00 % volume	14.42 Gy	✓
2	Evaluation dose (2 LO...	 Heart	At most 26.00 Gy dose at 2.00 % volume	14.66 Gy	✓
2	Plan dose: 2 LONG SB...	 Lungs-GTV	At most 20.00 Gy dose at 8.00 % volume	8.70 Gy	✓
2	Evaluation dose (2 LO...	 Lungs-GTV	At most 20.00 Gy dose at 8.00 % volume	9.57 Gy	✓
2	Plan dose: 2 LONG SB...	 SpinalCord_PRV05	At most 22.00 Gy dose at 2.00 % volume	1.01 Gy	✓
2	Evaluation dose (2 LO...	 SpinalCord_PRV05	At most 22.00 Gy dose at 2.00 % volume	1.04 Gy	✓
2	Plan dose: 2 LONG SB...	 Trachea	At most 32.00 Gy dose at 2.00 % volume	0.27 Gy	✓
2	Evaluation dose (2 LO...	 Trachea	At most 32.00 Gy dose at 2.00 % volume	0.30 Gy	✓

Figuur 104: Oplijsting van alle clinical goals (patiënt 4) en of ze al dan niet behaald zijn per Region of Interest. Plan dose is afkomstig van de originele structuren en evaluation dose is afkomstig van de teststructuren.

We merken op dat alle (afgeleide) teststructuren zijn geslaagd voor alle clinical goals. Daarbijkomend zijn de behaalde dosissen in de volumepercentages zeer gelijkend tussen de originele en de teststructuren.

## 4 Discussie

### 4.1 Pretrained ImageNet encoderweights

In tabel 6 zien we een significant verschil tussen de configuraties die wel voorgetrainde encodergewichten gebruiken (configuraties 1 & 2) en de configuraties die geen voorgetrainde encodergewichten (configuratie 3) gebruiken. Het is dus overduidelijk dat deze methoden het leerproces drastisch versnellen. Dit is in lijn met de resultaten van vorig onderzoek [5].

Daarnaast zien we geen significant verschil tussen de configuratie waar de encodergewichten wel getraind konden worden (configuratie 1) en de configuratie waar de encodergewichten niet getraind konden worden (configuratie 2). Dit ligt in lijn met onze verwachten dat algemeen de encodergewichten over een breed spectrum van doelen gebruikt kunnen worden en hier niet veel aan gesleuteld moet worden.

Ten laatste merken we graag op dat de testscores voor het getrainde orgaan (longen) in lijn liggen met de behaalde scores van andere papers (zie tabel 1) en de theoretische limiet (zie tabel 2).

### 4.2 Gewogen binaire crossentropie loss

Voor de longen bedroeg de test IoU score 0.881 bij de configuratie met drie patiënten en 0.916 bij de configuratie met 29 patiënten. De longen waren nagenoeg perfect ingetekend bij zowel de test- als bij de trainpatiënt. De WBCEL kan dus gebruikt worden om de longen te voorspellen. Dit is geen verrassing, maar wel een vereiste om een kans te hebben bij de overige organen.

Voor het hart bedroeg de test IoU score 0.467 bij de configuratie met drie patiënten. De intekening van het hart was nog niet compleet wenselijk bij de testpatiënt wegens relatief grote valse positieven. De voorspelling bij de trainpatiënt was niet wenselijk wegens relatief grote valse negatieven. Het niet voldoende intekenen van een orgaan kan leiden tot een te zware bestraling van dat orgaan. De test was echter op een te kleine dataset uitgevoerd om definitief een uitspraak te kunnen doen over het al dan niet bruikbaar zijn van de WBCEL bij het hart. Verdere testen met meer data zouden moeten bepalen of het model al dan niet bruikbaar is. Het model is alvast zeker in staat het orgaan min of meer terug te vinden.

Voor de slokdarm bedroeg de test IoU score 0.198 bij de configuratie met drie patiënten. Voor de slokdarm zijn er zowel bij de test- als bij de trainpatiënt nog relatief grote fouten. Het model is echter wel al in staat het orgaan kwalitatief terug te vinden. Testen op grotere groepen data zijn nodig om te bepalen of het model al dan niet bruikbaar is. Voor de kleine organen kunnen we algemeen vlugger een lage IoU score verwachten. Dit heeft als reden dat de meeste fouten van het model worden gemaakt aan de rand van het orgaan. Voor een klein orgaan maakt de rand een relatief groot deel uit van het totale oppervlak. Als gevolg zullen fouten bij kleine organen proportioneel zwaarder doorwegen dan bij grote organen.

Voor de luchtpijp bedroeg de test IoU score 0.771 bij de configuratie met drie patiënten. Zowel de test- als de trainpatiënt hebben nog relatief kleine fouten aan de rand van het orgaan. Deze wegen proportioneel zwaarder door bij kleine organen met als gevolg dat de test IoU misleidend laag lijkt vergeleken met de nagenoeg perfecte ingetekende organen. Dit model is zeker bruikbaar.



Voor het ruggenmerg bedroeg de test IoU score 0.345 bij de configuratie met drie patiënten. Zowel de test- als de trainpatiënt hebben relatief zeer veel valse positieven. Hierbij kan het voorspelde oppervlak makkelijk vier keer groter zijn dan het ware orgaan. Dit is echter minder erg dan valse negatieven en we kunnen concluderen dat het model op zijn minst het orgaan kan alloceren. Verdere testen op grotere groepen data zijn nodig om definitieve conclusies te trekken, maar we vermoeden dat dit zou zorgen voor een correctere uitlijning van de organen.

Voor het GTV bedroeg de test IoU score 0.013 bij de configuratie met drie patiënten waar enkel foto's werden gebruikt waar een GTV was op geannoteerd. De test IoU score bedroeg 0.001 indien we alle foto's van de patiënten gebruikten. De voorspellingen hadden zowel bij de test- als bij de trainpatiënten te veel valse positieven. Hoe het er nu naar uitziet is dit model zeker niet bruikbaar voor het GTV. Verdere testen zijn echter mogelijk waarbij we primair veel meer data nodig hebben en ook slim moeten omgaan met de gewichten voor de lossfunctie.

### 4.3 Gewogen categoriale crossentropie loss

Ondanks dat we een verbetering zagen in de maximale train IoU van 0.506 naar 0.546 tussen de configuratie met drie patiënten en de configuratie met 29 patiënten, zagen we geen verbetering bij de test IoU. Deze zakte van 0.587 naar 0.557. We hebben daarom de individuele test IoU scores bekeken van elke categorie. We zagen dat deze verbeterden voor het hart, het GTV en de longen, maar verslechterden voor de slokdarm, de luchtpijp en het ruggenmerg.

De longen waren nagenoeg perfect ingetekend. Deze kunnen dus parallel ingetekend worden met andere organen bij het gebruik van de WCCEL.

Het hart had nog relatief grote gebieden van valse positieven. Deze zijn echter minder erg dan valse negatieven. Het model is in staat het orgaan algemeen te lokaliseren, maar verdere verfijningen zijn nodig. Dit kan door enerzijds te testen op meer data en anderzijds de bepaling van de gewichten te herzien.

De slokdarm had relatief grote gebieden van valse positieven. Weer geldt de opmerking dat het model in staat is het orgaan te lokaliseren. De fouten zijn echter van een relatief groter kaliber in vergelijking met het hart, met relatieve fouten die drie keer de ware oppervlakte kunnen zijn. Ook hier zijn testen op meer data en een mogelijke herziening van de gewichten vereist.

De luchtpijp was nagenoeg perfect ingetekend met slechts minieme fouten aan de rand van het orgaan. We kunnen concluderen dat dit orgaan parallel ingetekend kan worden met andere organen bij het gebruik van de WCCEL.

Het ruggenmerg heeft kleine tot middelmatige relatieve fouten aan de rand van het orgaan. Dit zijn weer valse positieven, maar zijn beter dan bij de binaire situatie. Het intekenen is nog niet optimaal maar verdere verfijning lijkt ons mogelijk met meer data.

Het GTV is niet meer zo extreem verkeerd als bij het binaire geval. Het model is in staat het GTV terug te vinden maar maakt nog relatief grote valse positieve fouten. Het kaliber van deze intekeningsfouten lijken niet meteen overeen te komen met de lage test IoU scores van 0.074 (configuratie 1) tot 0.184 (configuratie 2). Het lijkt echter potentie te hebben en verdere testen zijn dan ook aangeraden.

## 4.4 Dicescore per orgaan

Voor de longen bedroeg de test Dicescore 0.960 bij de configuratie met drie patiënten. De longen waren nagenoeg perfect ingetekend bij zowel de test- als de trainpatiënt. De Dicescore is zeker bruikbaar bij het intekenen van dit orgaan. Dit is niet verbazend gezien de lage moeilijkheidsgraad, maar dit geeft een indicator over het mogelijk voorspelbaar zijn van de andere organen.

Voor het hart bedroeg de test Dicescore 0.876 bij de configuratie met drie patiënten. Het hart was nagenoeg perfect ingetekend bij de trainpatiënt, maar had nog kleine tot middelgrote fouten aan de rand bij de testpatiënt. Een deel hiervan bestond uit valse negatieven wat niet wenselijk is voor organen. We kunnen concluderen dat het model trainbaar is, maar dat meer data nodig is om dit orgaan beter in te tekenen.

Voor de slokdarm bedroeg de test Dicescore 0.477 bij de configuratie met drie patiënten. Deze lage score geeft echter een vertekend beeld als we kijken naar de intekeningen. De fouten doen zich voor aan de rand met als gevolg dat dit type fouten een relatief groot deel van de Dicescore beïnvloeden. Als we kijken naar de intekeningen zien we dat ze relatief goed overeenkomen met de realiteit. We kunnen concluderen dat dit model al relatief bruikbaar is maar dat verdere verbeteringen op uitlijning mogelijk zijn door meer data te gebruiken.

Voor de luchtpijp bedroeg de test Dicescore 0.856 bij de configuratie met drie patiënten. De organen zijn relatief goed ingetekend. Echter kan zich een dunne rand aan valse negatieven voordoen. Dit komt vermoedelijk doordat het model primair de luchtcaviteit in de luchtpijp herkent en niet nog het stuk weefsel daarrond. Dit kan echter verbeterd worden door meer data te gebruiken. We kunnen concluderen dat dit model goed bruikbaar is, maar dat verdere uitlijning van het weefsel rond de luchtcaviteit wenselijk is wat bereikt kan worden met meer trainingsdata.

Voor het ruggenmerg bedroeg de test Dicescore 0.798 bij de configuratie met drie patiënten. Opnieuw gaat de redenering op dat door de relatief kleine fouten aan de rand de test Dicescore verraderlijk klein lijkt. De fouten lijken zich namelijk makkelijk binnen de foutenmarge van de stralingsapparatuur te bevinden. We kunnen concluderen dat dit model bruikbaar is.

Voor het GTV bedroeg de test Dicescore 0.119 bij de configuratie met drie patiënten. Er werd geen tweede experiment uitgevoerd waar alle foto's werden gebruikt (ongeacht het al dan niet aanwezig zijn van een GTV). Dit omdat het eerste experiment geen extreem hoge FPR had. De reden van de lage test Dicescore was het zeer overtraint zijn van het model aangezien de maximale train Dicescore 0.735 bedroeg. De test Dicescore lijkt echter niet beter te krijgen omdat de validatie Dicescore ook maar zo lage waarden haalde. Het ingetekende GTV bij de testpatiënt was veel te klein ten opzichte van de realiteit. Dit is niet wenselijk binnen de radiotherapie omdat een stralingsbehandeling een veel te kleine "coverage" zou hebben. Het is echter wel interessant dat het model überhaupt in staat was een stukje van het GTV te herkennen. In dat opzicht heeft dit model potentieel om een zeker type van outlier detectie te hebben. Verdere testen zijn nodig met mogelijks een andere metriek, waar een score wordt toegeschreven of het model al dan niet iets herkend van het GTV.

## 4.5 Dixeloss op alle organen

De test Dicescore had een verbetering van 0.616 naar 0.677 van de configuratie met drie patiënten naar de configuratie met 29 patiënten. We haalden aan dat deze score een vertekend beeld kan geven. Een pixel die 0.99 kans heeft om long te zijn en 0.4 kans om hart te zijn, zou aanzien moeten worden als long en niet de onterechte straf van de 0.4 kans hart mogen krijgen. Met die redenering hebben we de Dicescores berekend voor alle organen met verschillende thresholds voor de kansen. We zagen dat er voor alle organen een verbetering was gaande van configuratie 1 naar configuratie 2. We zagen ook dat de Dicescores tot op hoge thresholds relatief stabiel bleven.

De longen hadden bij configuratie 2 een test Dicescore van 0.976 bij een threshold van 0.5. Ze waren nagenoeg perfect ingetekend en we kunnen concluderen dat ze parallel kunnen ingetekend worden met andere organen bij het gebruik van de Dixeloss.

Het hart had bij configuratie 2 een test Dicescore van 0.853 bij een threshold van 0.5. Ze was nagenoeg perfect ingetekend. Er waren slechts relatief kleine fouten aan de randen die sterk verbeterd waren van configuratie 1 naar configuratie 2 en waar we dus kunnen voorspellen dat ze eventueel nog beter kunnen worden bij het trainen op nog meer data. We kunnen concluderen dat het hart parallel ingetekend kan worden met andere organen bij het gebruik van de Dixeloss.

De slokdarm had bij configuratie 2 een test Dicescore van 0.630 bij een threshold van 0.5. Deze fouten zijn afkomstig van de relatief middelmatige fouten aan de rand van het orgaan. Deze fouten zijn echter verbeterd gaande van configuratie 1 naar configuratie 2 en we verwachten dat dit verder verbetert bij het trainen op nog meer data. We raden een training aan op nog meer data om definitief een uitspraak te kunnen doen over het intekenen van de slokdarm parallel aan de andere organen.

De luchtpijp had bij configuratie 2 een test Dicescore van 0.808 bij een threshold van 0.5. De fouten liggen weer aan het niet compleet intekenen van het weefsel rondom de luchtcaviteit. Dit lijkt verder verbeterbaar bij het trainen op meer data. Het model is echter relatief goed in staat het orgaan te herkennen aangezien de rand van valse negatieven redelijk parallel loopt met het ware orgaan. We raden een verdere training aan in de hoop deze foutieve rand verder uit te lijnen. Het model lijkt echter nu al trainbaar op de luchtpijp parallel aan de andere organen.

Het ruggenmerg had bij configuratie 2 een test Dicescore van 0.784 bij een threshold van 0.5. Als we kijken naar de foto's zien we dat deze score verraderlijk laag ligt. De fouten zijn relatief klein ten opzichte van het orgaan en zullen binnen de foutenmarge van de bestralingsapparatuur liggen. We kunnen concluderen dat het ruggenmerg parallel ingetekend kan worden met andere organen bij het gebruik van de Dixeloss.

Het GTV had bij configuratie 2 een test Dicescore van 0.086 bij een threshold van 0.5. Op de testfoto werd het compleet niet herkend. Vermoedelijk is het model zeer goed geworden in het veralgemeniseren waardoor kleine variaties zoals het GTV niet direct gemarkeerd worden. We zien hier niet meteen een mogelijkheid tot verbetering in.

## 4.6 DHV en clinical goals

De longen waren nagenoeg perfect ingetekend. Slechts een drietal foto's bij één patiënt weken af van de realiteit omdat het model enkele luchtcaviteiten in de darmen aanschouwde als stukken long. Betreffende de andere klinische metrieken kwamen de longen zo goed als perfect overeen. De longen slagen dus voor de klinische testen en zouden gebruikt kunnen worden in de praktijk.

Het hart week af bij het begin en het einde van het orgaan. Omdat deze afwijkingen, althans gering in aantal, groot waren ten opzichte van de realiteit en het hart op relatief weinig slices aanwezig is, kon dit resulteren in middelmatige tot grote afwijkingen in het ingetekend volume. Echter uitte deze afwijkingen zich niet vaak tot grote afwijkingen in het DVH en de clinical goals omdat het ontbreken van de intekening op slechts een paar slices geen groot globaal effect had. Het hart slaagt bijgevolg voor de meeste klinische testen, maar er moet gekeken worden deze discrepanties aan de rand verder te verfijnen. Wij oordelen dat het gebruikt zou kunnen worden in de praktijk.

De slokdarm had veel regio's waar er grote afwijkingen waren van de realiteit. Dit uitte zich vaak in middelmatige tot grote afwijkingen in de ingetekende volumes. Ook bij de DVH's waren er vaak afwijkingen. Het probleem van de slokdarm ligt in het zeer variabel zijn van het orgaan. Het volgt geen recht pad zoals bijvoorbeeld het ruggenmerg of de luchtpijp. Daarnaast kan het sterk variëren in grootte waardoor het model de grootte niet gemakkelijk kan inschatten. Ondanks deze afwijkingen oordeelde de radiotherapeut-oncoloog dat het orgaan globaal goed was ingetekend. Wij raden verdere trainingen met grote datasets aan met voldoende verschillende foto's om zo alsnog betere intekeningen te verkrijgen.

De luchtpijp was goed ingetekend met slechts afwijkingen van een slice of twee bij het begin en/of einde van het orgaan. Dit komt vermoedelijk door de klinische definities over de overgang van keel naar luchtpijp en van luchtpijp naar bronchi. De relatief grote afwijkingen bij de ingetekende volumes zijn hier deels het gevolg van en deels doordat kleine niet-relevantie variaties aan de rand van het orgaan een relatief grote impact hebben wegens de relatief kleine doorsneden. Algemeen slaagt de luchtpijp voor de klinische testen en zou het gebruikt kunnen worden in de praktijk.

Het ruggenmerg was goed ingetekend met af en toe een serie van rond de drie CT-beelden waar het model het orgaan niet had ingetekend. Ook bij de uiteinden van het orgaan was er soms discussie tot waar het orgaan moest ingetekend worden. Deze randintekeningen zijn echter niet relevant aangezien ze niet mee het stralingsplan zullen bepalen. Opnieuw geldt de opmerking dat de ingetekende volumes sterk konden afwijken, maar dat dit het gevolg is van niet-relevante afwijkingen aan de rand van het orgaan. Algemeen slaagt het ruggenmerg voor de klinische testen en zou het gebruikt kunnen worden in de praktijk.

## 4.7 Vergelijking vorige onderzoeken

We merkten dat onze resultaten niet geheel overeen kwamen met resultaten van andere onderzoeken en hier regelmatig onder lagen. Indien we echter onze resultaten vergeleken met de uitgemiddelde resultaten van een grote groep radiotherapeuten-oncologen zagen we dat we nagenoeg dezelfde scores haalden. Vermoedelijk waren de datasets van de aangehaalde onderzoeken niet groot en/of variabel genoeg. We concluderen hieruit dat onze resultaten competitief zijn met de resultaten van de clinici.

## 4.8 Klinische metrieken

Het gebruik van klinische metrieken zoals de DHV's, ROI volumes en clinical goals bewees hun nut in het meten van het effect van verschillende type afwijkingen. Het leert ons ook of de afwijkingen van ons model met de realiteit klinisch relevant zijn of niet. Zo zagen we dat de fouten aan de randen van de seriële organen niet bepaald relevant waren. Deze fouten konden verraderlijke lage Dicescores geven, maar gaven geen klinisch verschil. Het is dan ook aan te raden dat verdere onderzoeken naar autosegmentatie in CT-beelden zeker gebruik maken van deze metrieken en hun segmentaties voorleggen aan radiotherapeuten-oncologen.

## 4.9 Verder onderzoek

Het was niet mogelijk het GTV te leren herkennen. Er leek soms hoop te zijn dat het model op zijn minst een klein deel van het GTV kon voorspellen. Echter leek dit niet iets consistent te zijn. Het is duidelijk dat het leren herkennen van tumoren een andere aanpak vereist dan onze methoden. Daartegenover merkten we dat het model in staat was bijna alle organen tot klinisch relevant niveau in te tekenen. Voor de slokdarm zien wij nog verbeteringen door in eerste instantie te trainen op meer data. We hebben echter slechts één derde van onze data gebruikt en nog niet alle organen beschouwd zoals de bronchi, de aorta,... De volgende logische stap lijkt ons om het model verder te verfijnen door te trainen op grotere datasets en uit te breiden door nog meer organen te leren herkennen. We raden echter aan dat alle data vooraf goed bekeken wordt opdat er geen inconsistenties zijn die het model kunnen verwarren. We hebben het daarbij bijvoorbeeld over het standaardiseren tot waar de seriële organen worden ingetekend en hoe breed ze worden ingetekend.

Bij het bekijken van de voorspelde organen in Raystation merkten we soms kleine secties waar het model sterk afweek van de realiteit. Het is bijvoorbeeld niet logisch dat er plots drie slices geen ruggenmerg zou zijn. Ook waren de ingetekende overgangen van hart naar niet-hart vaak zeer drastisch. Wij stellen voor om een nieuw model te trainen dat gebruikt maakt van lokale 3D-informatie. Concreet moet het model getraind worden op het intekenen van de organen aan de hand van niet alleen het huidige CT-beeld, maar ook één à twee slices ervoor en erna. Deze extra contextuele informatie zal er volgens ons voor zorgen dat er geen of alleszins minder abrupte of onlogische overgangen zullen zijn bij de ingetekende organen.

## 5 Conclusie

Het is zeker de moeite waard om voorgetrainde encodergewichten te gebruiken bij het trainen van segmentatiemodellen. Wij raden dan ook iedereen aan hier gebruik van te maken.

Het gebruik van een gewogen binaire crossentropie loss voor het segmenteren van de organen leek veelbelovend tot bruikbaar, behalve voor het GTV. Verdere testen op grotere datasets zijn nodig in de hoop dezelfde evolutie te zien als bij de Dixeloss van een configuratie met weinig naar een configuratie met veel patiënten.

Het gebruik van een gewogen categoriale crossentropie loss is de logische extensie indien het binair model goed werkt. Het model lijkt veelbelovend, maar meer testen en eventueel een herziening van de gewichten is nodig. Opnieuw is een zeer andere aanpak vereist voor het GTV. Het model is wel in staat op zijn minst het orgaan terug te vinden. Dit geeft hoop tot verdere verfijningen. Het model kan alleszins gebruikt worden om een ruwe intekening van de organen te maken. Ook dit vergemakkelijkt het werk van de radiotherapeuten-oncologen die dan manueel de contouren verder kunnen verfijnen.

De Dixeloss was zeer goed in staat de organen in te tekenen. Dit was zichtbaar bij de Dicescores die van dezelfde orde waren als de Dicescores tussen intekeningen van verschillende clinici. Bij het GTV liet het model het weer afweten. De klinische metrieken bewezen dat de resterende fouten in de organen vaak niet van een relevante orde of aard waren en dat het model globaal bruikbaar is voor autosegmentatie met af en toe een afwijking. Het controleren van de segmentaties door een radiotherapeut-oncoloog is een mogelijke oplossing, maar algemeen vergemakkelijkt het model drastisch het werk.

De klinische metrieken bewezen hun nut bij het oordelen of het model al dan niet nuttig was. Het is een zeer belangrijke toets achteraf omdat het effect in de praktijk het belangrijkste blijft van eenderwelke segmentatie. Wij raden dan ook iedereen aan hier gebruik van te maken.

Gezien de globaal uitstekende resultaten is het aangewezen het model verder te verfijnen op meer data en uit te breiden op meer organen. Daarnaast lijkt een aanpak waar enkele lokale slices worden gebruikt voor het segmenteren van één slice ons uitermate veelbelovend in het wegwerken van de meest voorkomende fouten die zich nu nog voordoen.

## 6 Dankwoord

Deze thesis zou niet mogelijk zonder een hoop mensen. Graag zou ik deze willen bedanken via deze niet-exhaustieve oplijsting.

Zonder twijfel verdient mijn promotor Prof. Dr. ir. Barbara Vanderstraeten het meeste dank. Wat ik wou in mijn thesis was duidelijk ,maar algemeen: “iets met radioactiviteit/-therapie, en er mag gerust een hoop programmeren zijn.” Ze leverde me enkele ideeën, waaronder het idee van deze thesis dat nog niet was uitgewerkt op onze universiteit. Een dergelijke stap in het duister leek mij wel iets hebben. Je hebt een groter aspect van avontuur en vrijheid. Achteraf heb ik hard genoten van deze aspecten en ben ik tevreden met het resultaat. Uiteraard had ik graag nog meer gedaan, maar ik ben zeker dat Barbara in staat zal zijn dezelfde interesse op te wekken bij een volgende student(e) als ze bij mij heeft kunnen realiseren en zo het werk verder te zetten. Deze student(e) zal ook zeker kunnen rekenen op de hulpvaardigheid en snelle respons van Barbara, alsook delen in de interesse en nieuwsgierigheid die ook zij had en zal hebben.

Als copromotor stond Prof. Dr. Luc Van Hoorebeke in voor het administratief aspect van de thesis. Hij zorgde ervoor dat we in orde waren met zaken zoals het houden van een tussentijdse evaluatie, het regelen van de finale presentatie en het vervullen van de modaliteiten.

Ongeacht de grote bronnen van informatie die men kan vinden op het internet, heb je nog steeds deskundigen nodig. Mensen die jou kunnen helpen met je problemen, je vragen en je verdere ideeën kunnen geven en kunnen adviseren. Hierin zijn Dr. ir. Jan Aelterman en Dr. Eva Vandersmissen meer dan geslaagd. Ze waren altijd bereid me te helpen met problemen en deskundig advies te verschaffen. Ze gaven me inzichten tot mogelijke methoden, verbeteringen en experimenten en zonder hen was deze thesis vermoedelijk maar half zo lang.

Graag bedank ik ook de thoracale radiotherapeut-oncologen en dosimetristen voor het leveren van grote hoeveelheden geanonimiseerde data en in het bijzonder dosimetrist Bruno Goddeeris voor de export vanuit RayStation en zijn expert opinies van de resultaten van het model.

Sinds de “lockdown” had mijn vriendin Clara Tanghe het (on)genoegen om 24/7 met mij samen te leven. Ze kon delen in mijn momenten van euforie en frustraties, en heeft daarnaast een voortreffende rol als “rubber ducky” (die ook nog eens kon terugpraten) ingenomen. Daarnaast heeft ze tot in de kleinste detail mijn thesis nagelezen, waarvoor ik vermoedelijk volgend jaar de prijs ga moeten betalen wanneer zij haar thesis maakt.

Het onderwerp van mijn thesis sprak ook enkele vriend(inn)en aan, die graag af en toe een kijkje namen en het uiteraard niet konden laten mijn taal te verbeteren. Hiervoor bedank ik Nathan Steyaert, Justine D’Hoine, Ellen Van de Steen en Tessa Ickx.

Mijn familie heeft me altijd gesteund in mijn studies en is onwaarschijnlijk trots op mij voor het (hopelijk) behalen van mijn Master of Science in de fysica en de sterrenkunde.

Ik begon deze lijst met het zeggen dat ik een hoop mensen wou bedanken. Ik was niet helemaal eerlijk aangezien ik de lijst eindig met de HPC. We werken altijd maar meer en grotere datasets, met complexere programma’s en hebben hierbij steeds zwaardere hardware nodig. De HPC zorgt ervoor dat enorm veel mensen aan de slag kunnen met hardware die ze vermoedelijk zelf nooit zouden kunnen betalen.

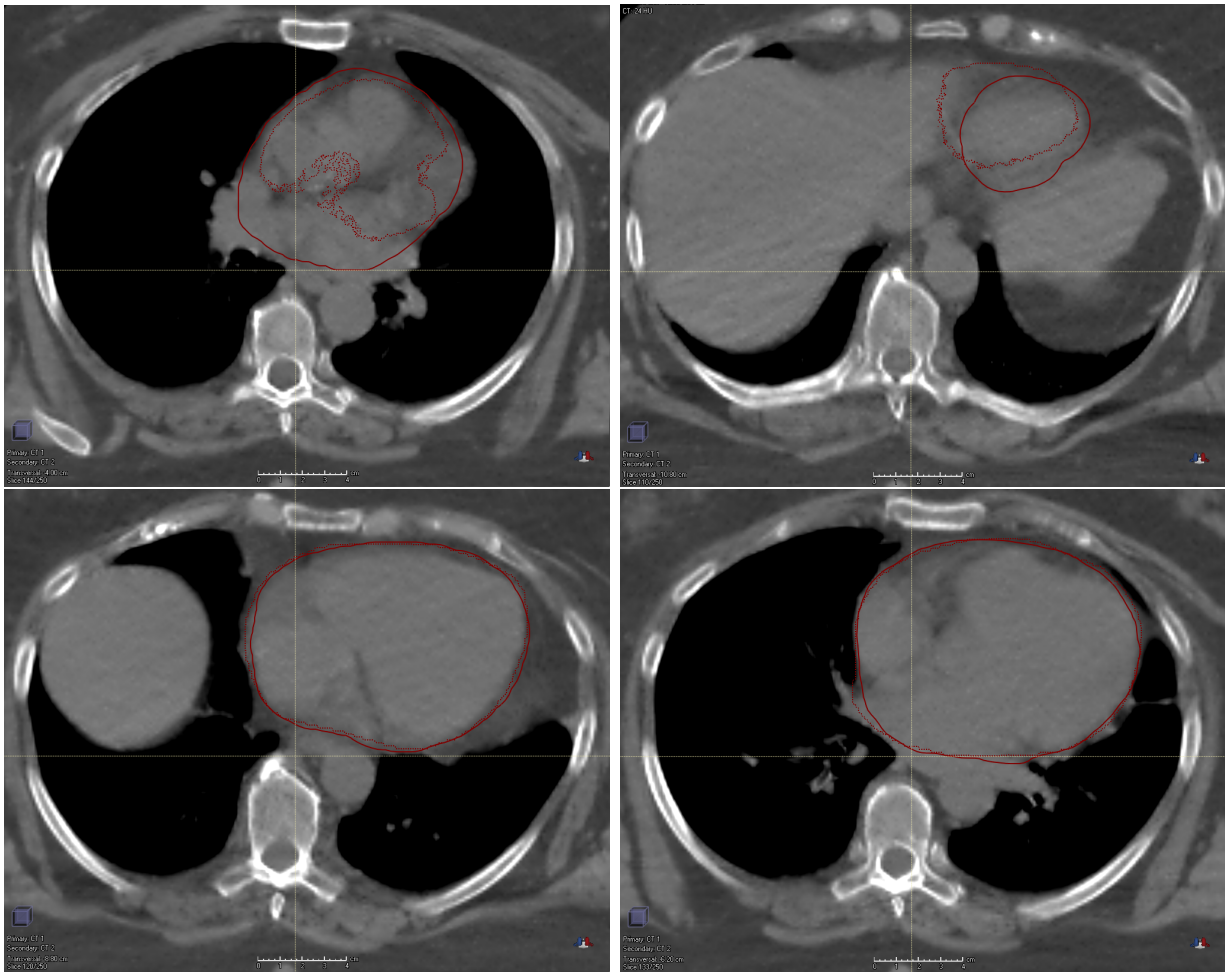
## Referenties

- [1] Statistiekvlaanderen.be. 2020. *Doodsoorzaken* [online] Geraadpleegd van <https://www.statistiekvlaanderen.be/nl/doodsoorzaken>.
- [2] Salvat, F. and Fernández-Varea, J., 2009. Overview of physical interaction models for photon and electron transport used in Monte Carlo codes. *Metrologia*, 46(2), pp.S112-S138.
- [3] Davisson, C. M. (1965). "Interaction of gamma-radiation with matter". In Kai Siegbahn (ed.). *Alpha-, Beta- and Gamma-ray Spectroscopy: Volume 1. Alpha-*. 1. Amsterdam: North-Holland Publishing Company. pp. 37–78.
- [4] Ronneberger, O., Fischer, P. and Brox, T., 2020. *U-Net: Convolutional Networks For Biomedical Image Segmentation*. [online] arXiv.org. Geraadpleegd van <https://arxiv.org/abs/1505.04597>.
- [5] Zhou, X., Takayama, R., Wang, S., Hara, T. and Fujita, H. (2017). Deep learning of the sectional appearances of 3D CT images for anatomical structure segmentation based on an FCN voting method. *Medical Physics*, 44(10), pp.5221-5233.
- [6] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. *Proc CVPR*. 2015;3431–3440.
- [7] Vesal, S., Ravikumar, N. and Maier, A., 2020. *A 2D Dilated Residual U-Net For Multi-Organ Segmentation In Thoracic CT*. [online] arXiv.org. Geraadpleegd van <https://arxiv.org/abs/1905.07710>.
- [8] Dong, X., Lei, Y., Wang, T., Thomas, M., Tang, L., Curran, W., Liu, T. and Yang, X., 2020. *Automatic Multiorgan Segmentation In Thoraxctimages Using U-Net-GAN*.
- [9] Tsang, Y., Hoskin, P., Spezi, E., Landau, D., Lester, J., Miles, E. and Conibear, J., 2019. Assessment of contour variability in target volumes and organs at risk in lung cancer radiotherapy. *Technical Innovations Patient Support in Radiation Oncology*, 10, pp.8-12.
- [10] www-pub.iaea.org. 2020. [online] Geraadpleegd van [https://www-pub.iaea.org/MTCD/Publications/PDF/P1679\\_HH31\\_web.pdf](https://www-pub.iaea.org/MTCD/Publications/PDF/P1679_HH31_web.pdf).
- [11] En.wikipedia.org. 2020. *Backpropagation*. [online] Geraadpleegd van <https://en.wikipedia.org/wiki/Backpropagation>.
- [12] de Vos, B., Wolterink, J., de Jong, P., Viergever, M. and Išgum, I., 2020. *2D Image Classification For 3D Anatomy Localization: Employing Deep Convolutional Neural Networks*. <https://www.spiedigitallibrary.org/>.
- [13] Trullo, R., Petitjean, C., Nie, D., Shen, D. and Ruan, S., 2020. *Joint Segmentation Of Multiple Thoracic Organs In CT Images With Two Collaborative Deep Architectures*. [online] Geraadpleegd van <https://www.researchgate.net/publication/319640486>.
- [14] Vesal, S., Ravikumar, N. and Maier, A., 2020. *A 2D Dilated Residual U-Net For Multi-Organ Segmentation In Thoracic CT*. [online] arXiv.org. Geraadpleegd van <https://arxiv.org/abs/1905.07710>.

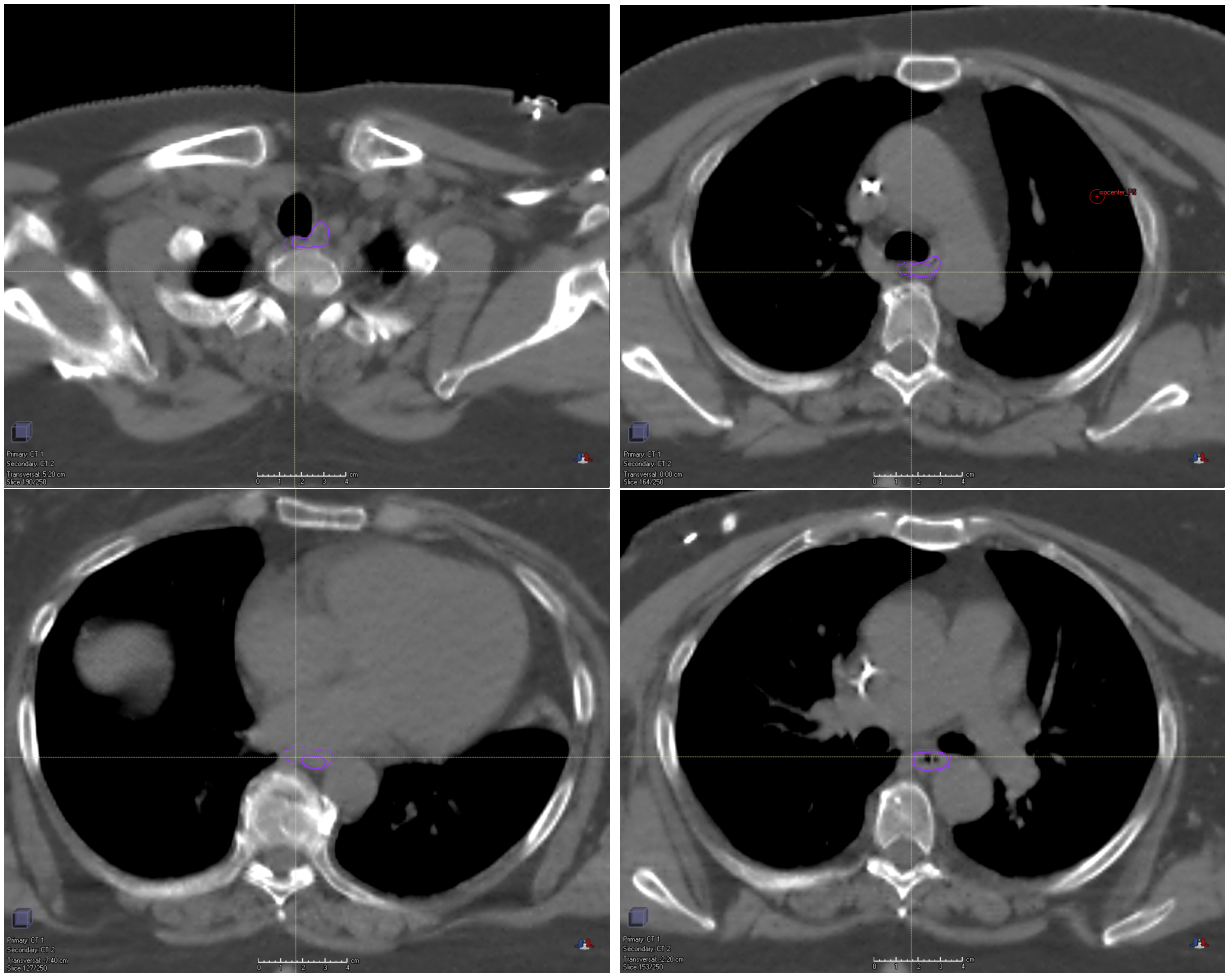


- [15] Tiu, E., 2020. *Metrics To Evaluate Your Semantic Segmentation Model*. [online] Medium. Geraadpleegd van <https://towardsdatascience.com/metrics-to-evaluate-your-semantic-segmentation-model-6bcb99639aa2>.

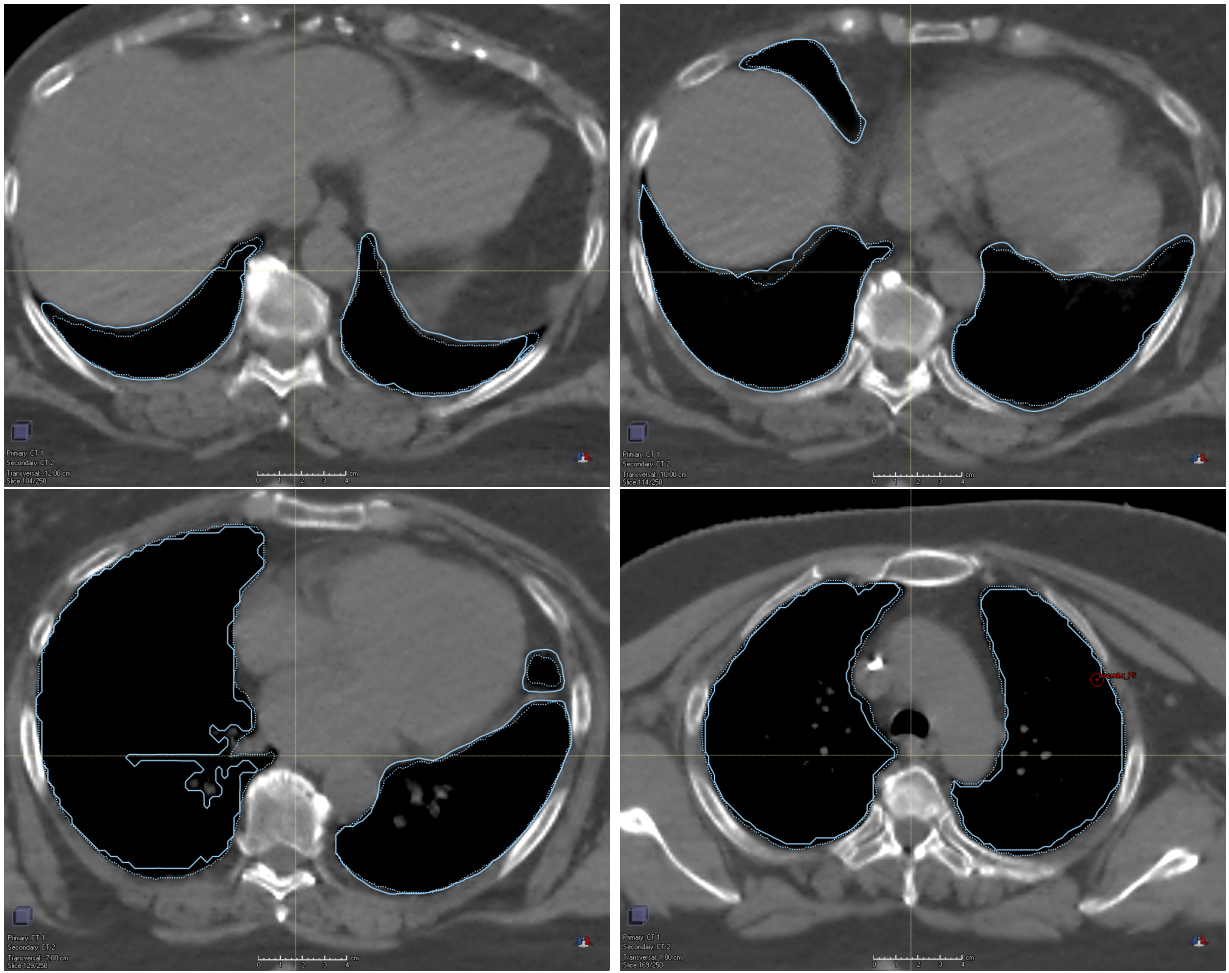
## A Vergelijking originele en teststructuren in RayStation



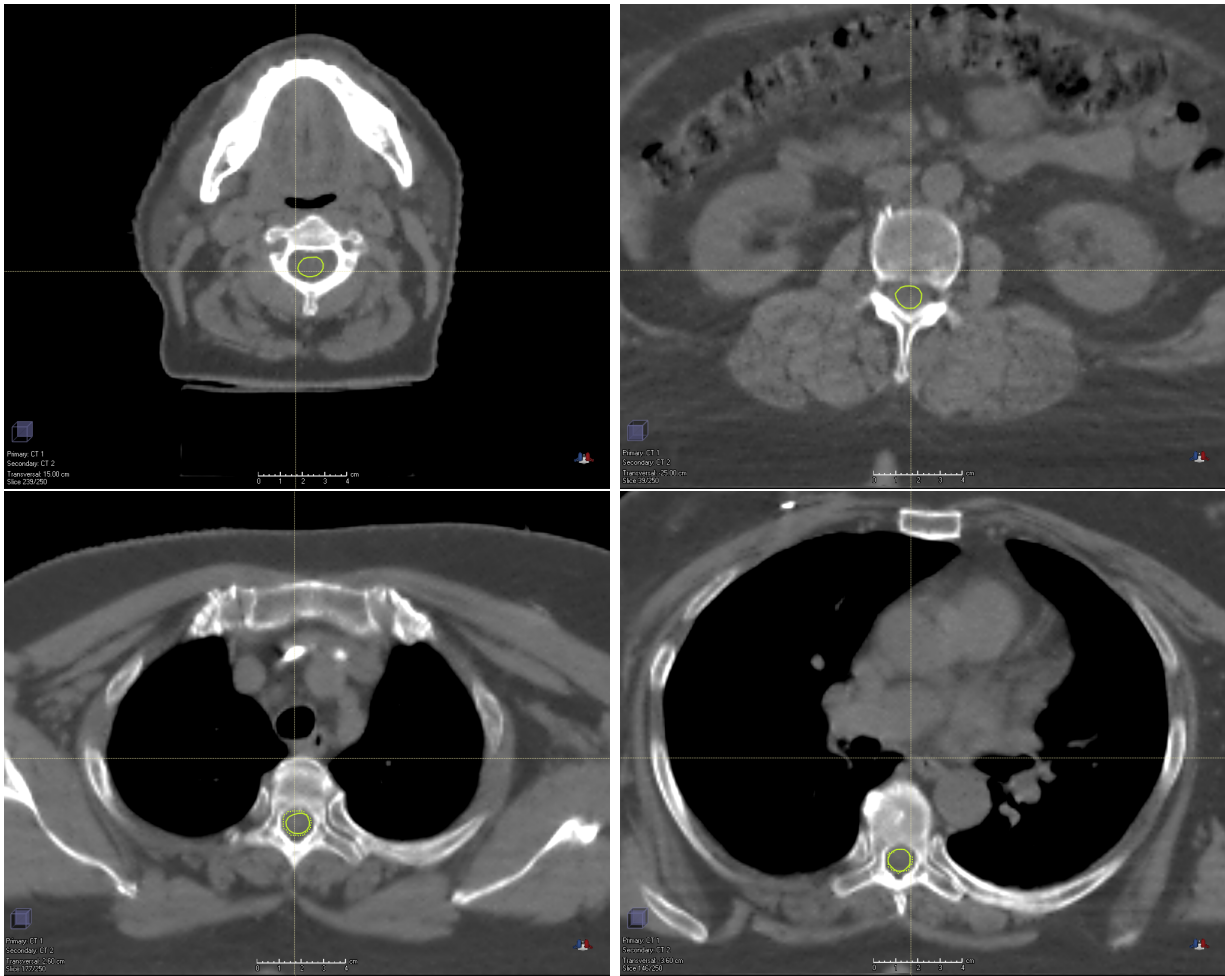
Figuur 105: Illustratie waar het verschil zeer sterk is tussen het originele en testhart (testpatiënt één) bij het begin van het orgaan (linksboven) en bij het einde van het orgaan (rechtsboven). Willekeurige beelden waar de intekeningen in orde waren (onder). Volle lijn: originele structuren, stippellijn: teststructuren.



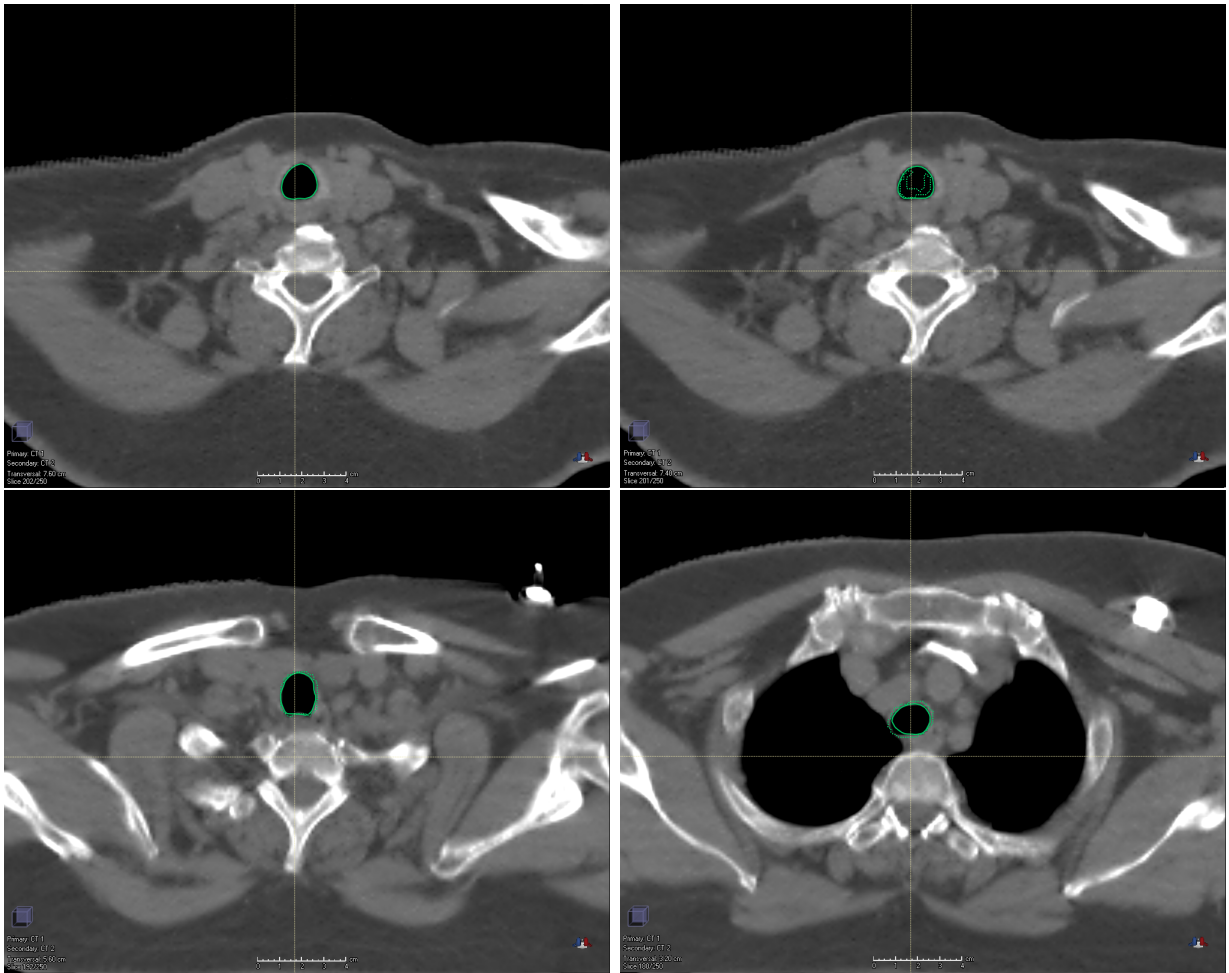
Figuur 106: Illustratie waar het verschil zeer sterk is tussen de originele en testslokdarm (test-patiënt één) in serie één (linksboven), serie twee (rechtsboven), serie drie (linksonder). Wilkeurig beeld waar de intekening in orde was (rechtsonder). Volle lijn: originele structuren, stippellijn: teststructuren.



Figuur 107: Illustratie waar de intekeningen in orde waren voor de longen (testpatiënt één). Volle lijn: originele structuren, stippellijn: teststructuren.

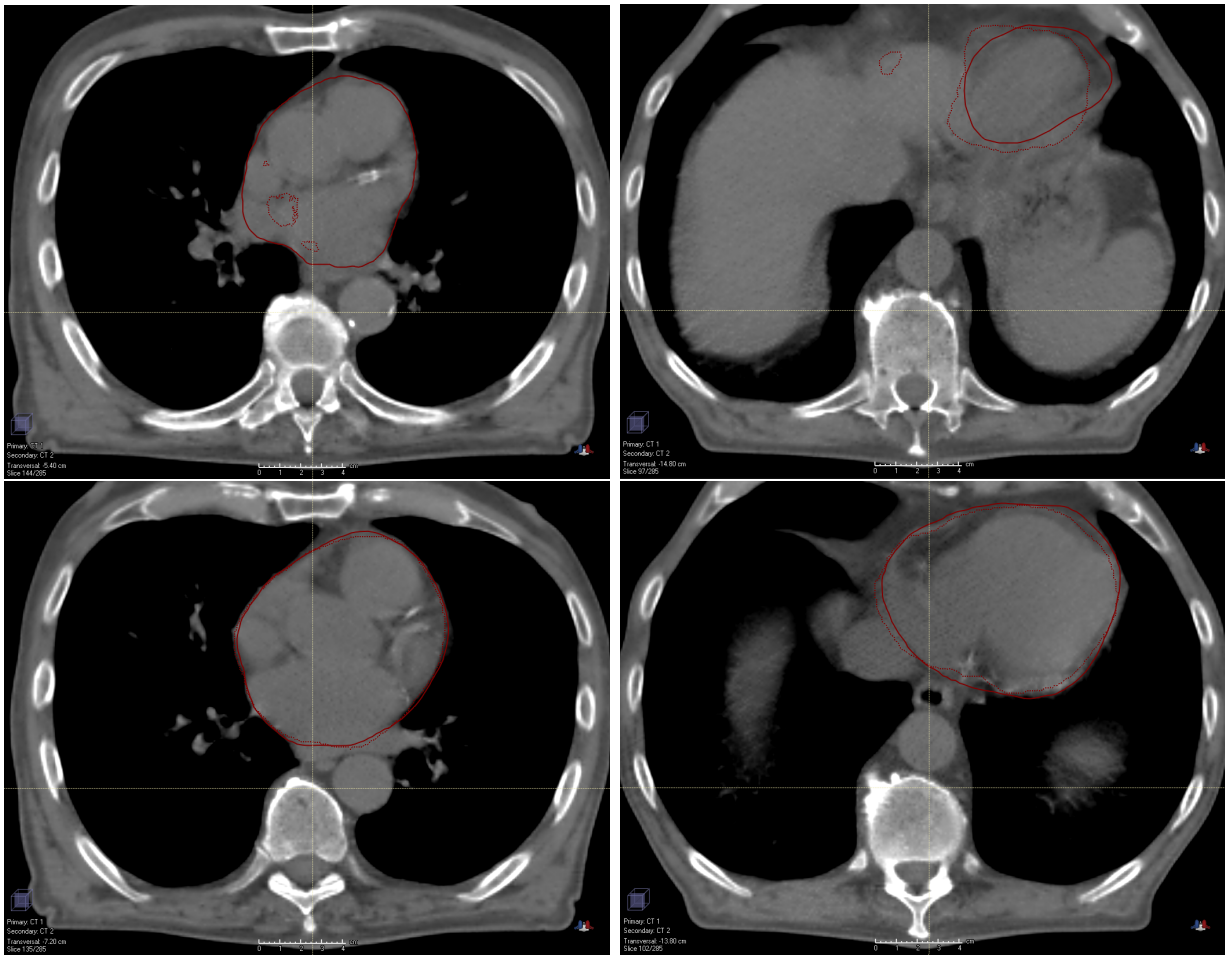


Figuur 108: Illustratie waar het intekenen niet in orde was voor het ruggenmerg (testpatiënt één) bij het begin van het orgaan (linksboven) en het einde van het orgaan (rechtsboven). Wilkekeurige beelden waar de intekeningen in orde waren (onder). Volle lijn: originele structuren, stippellijn: teststructuren.

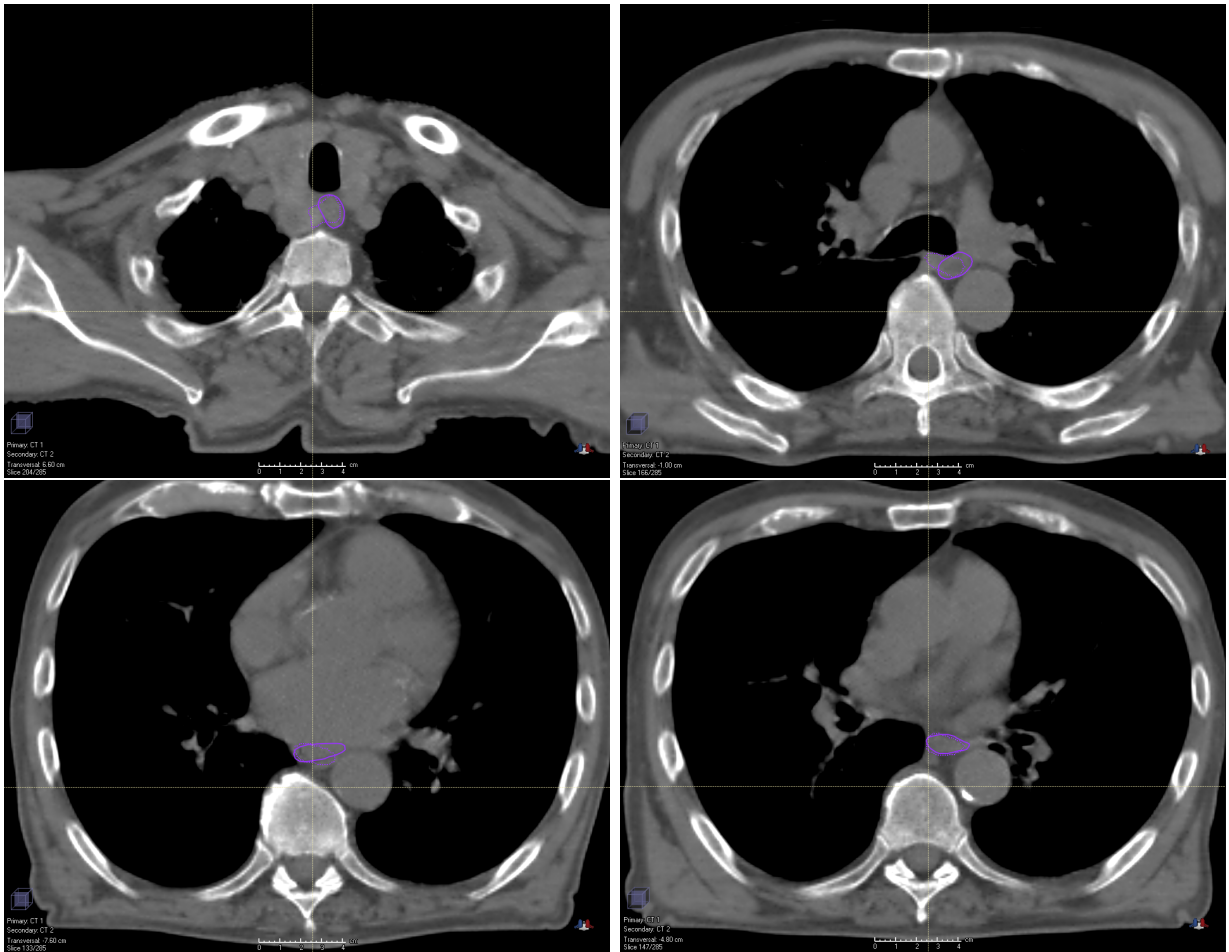


Figuur 109: Illustratie waar het intekenen niet in orde was voor de luchtpijp (testpatiënt één) bij intekening één bij het begin van het orgaan (linksboven) en bij intekening twee bij het begin van het orgaan (rechtsboven). Willekeurige beelden waar de intekeningen in orde waren (onder). Volle lijn: originele structuren, stippellijn: teststructuren.



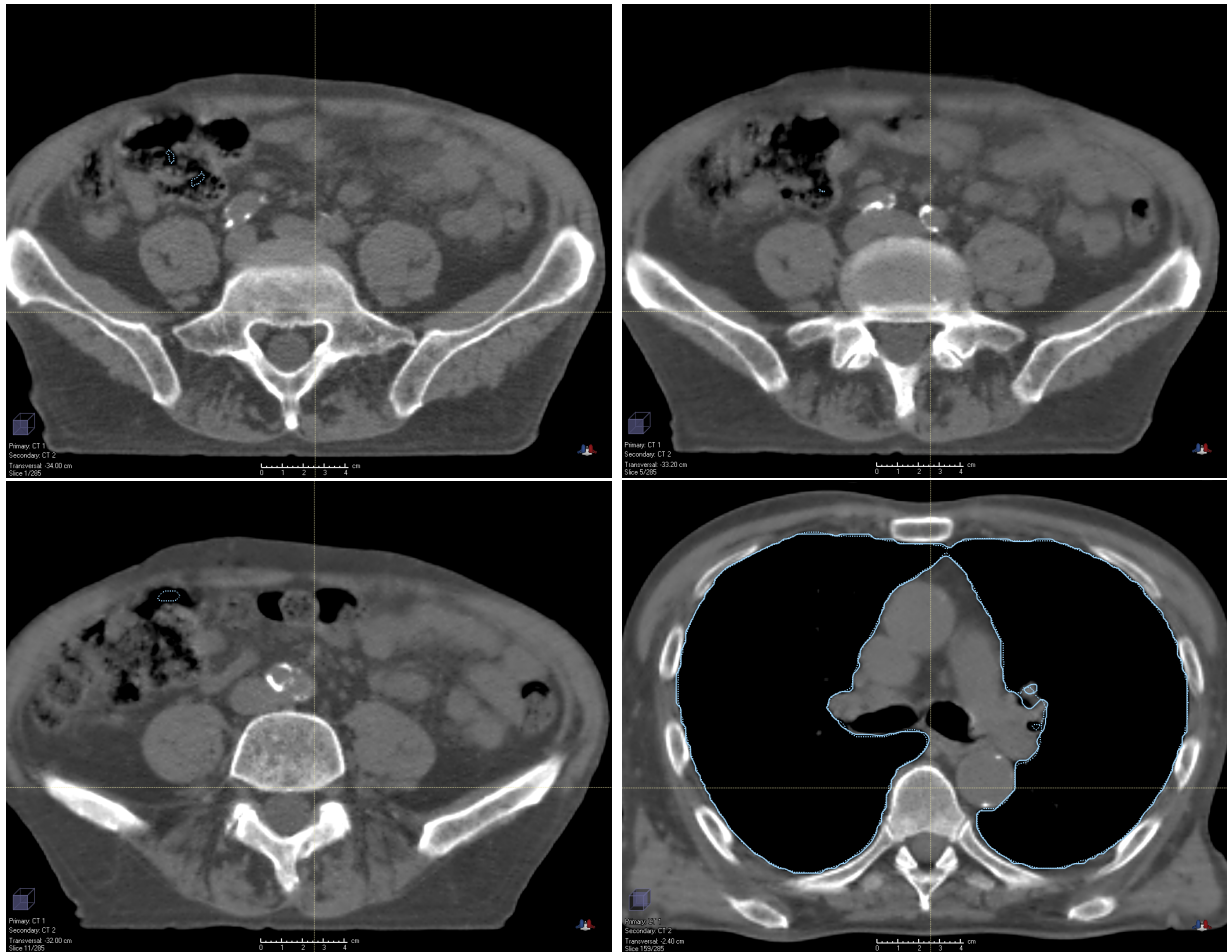


Figuur 110: Illustratie waar het intekenen niet in orde was voor het hart (testpatiënt twee) bij het begin van het orgaan (linksboven) en bij het einde van het orgaan (rechtsboven). Willekeurige beelden waar de intekening in orde was (onder). Volle lijn: originele structuren, stippellijn: teststructuren.

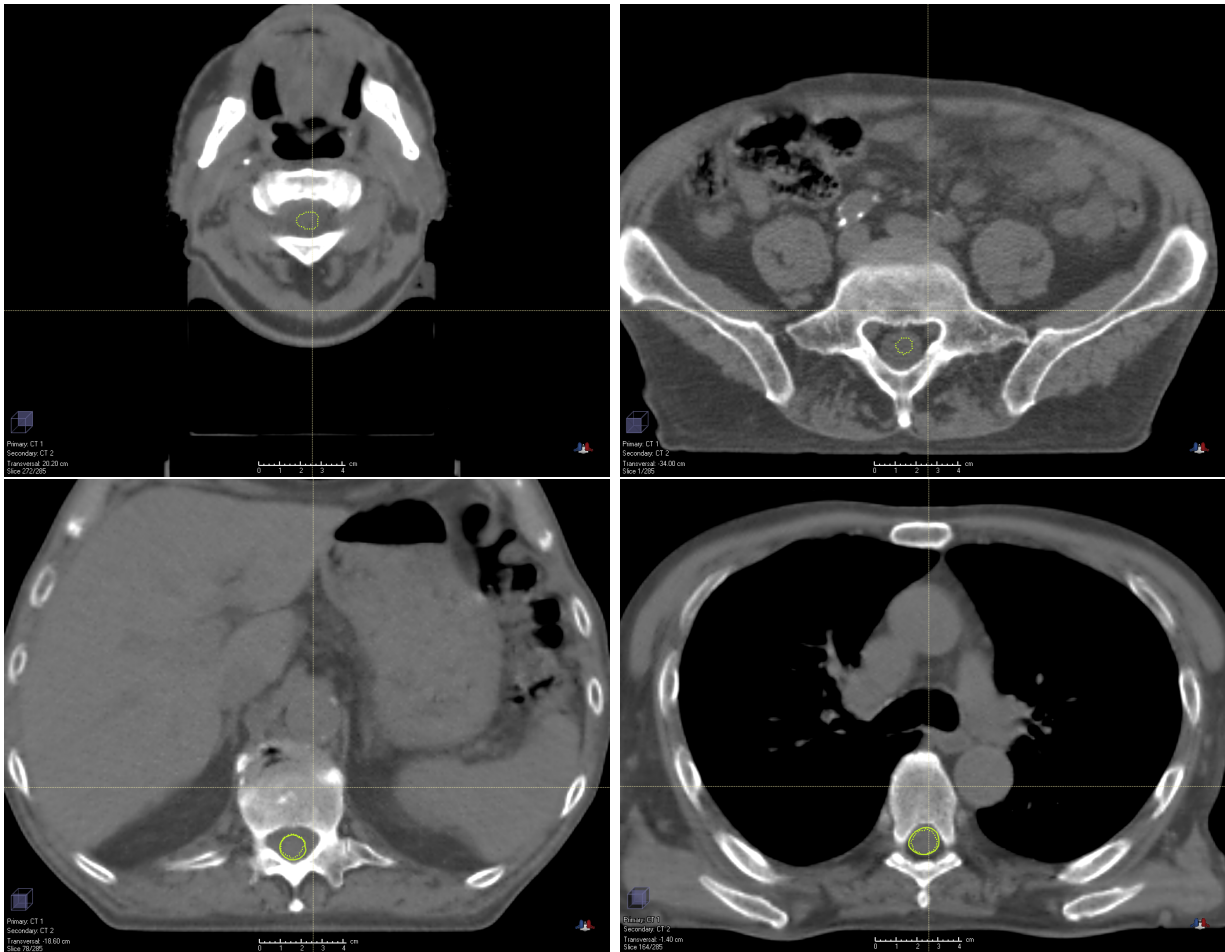


Figuur 111: Illustratie waar het intekenen niet in orde was voor de slokdarm (testpatiënt twee) in serie één (linksboven), serie twee (rechtsboven), serie drie (linksonder). Willekeurig beeld waar de intekening in orde was (rechtsonder). Volle lijn: originele structuren, stippellijn: teststructuren.

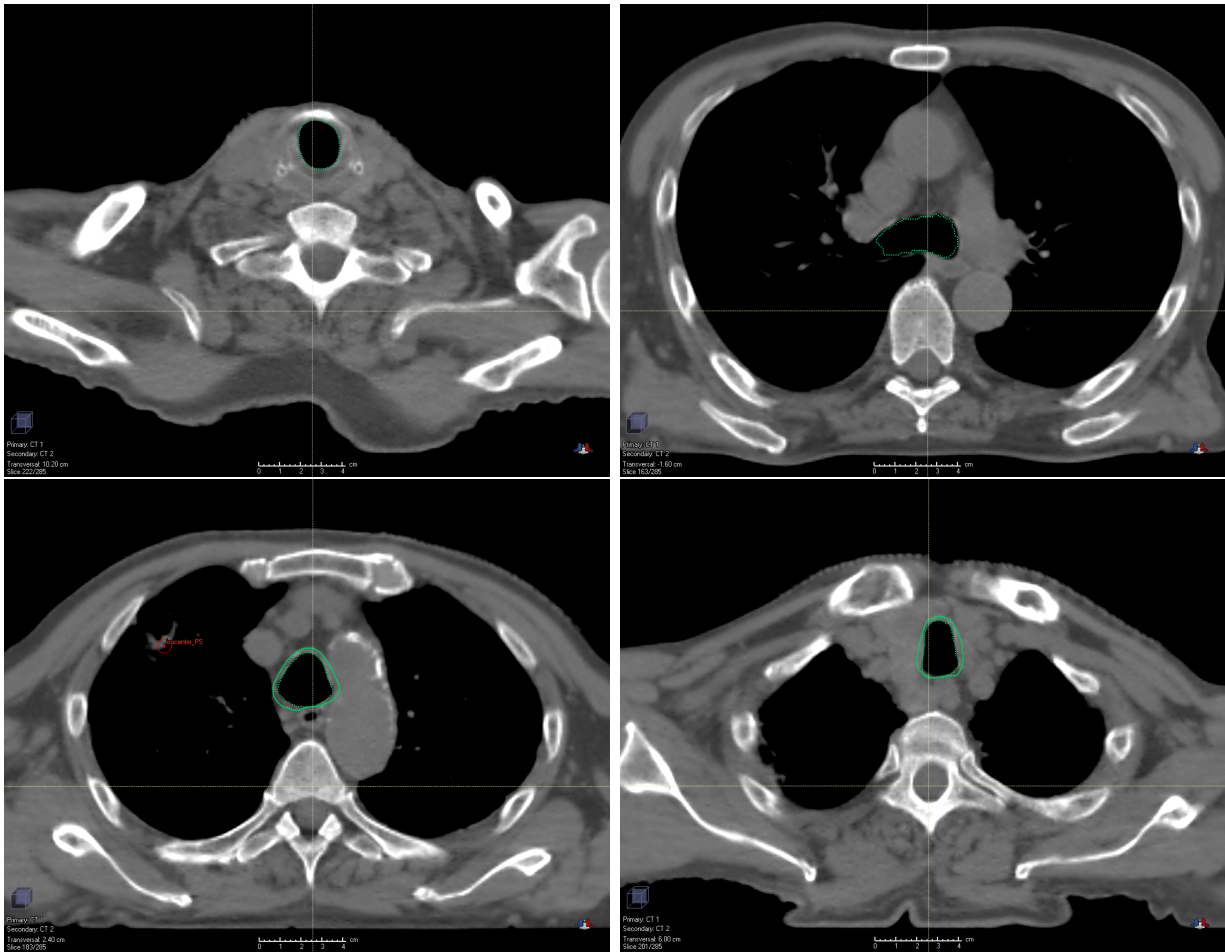




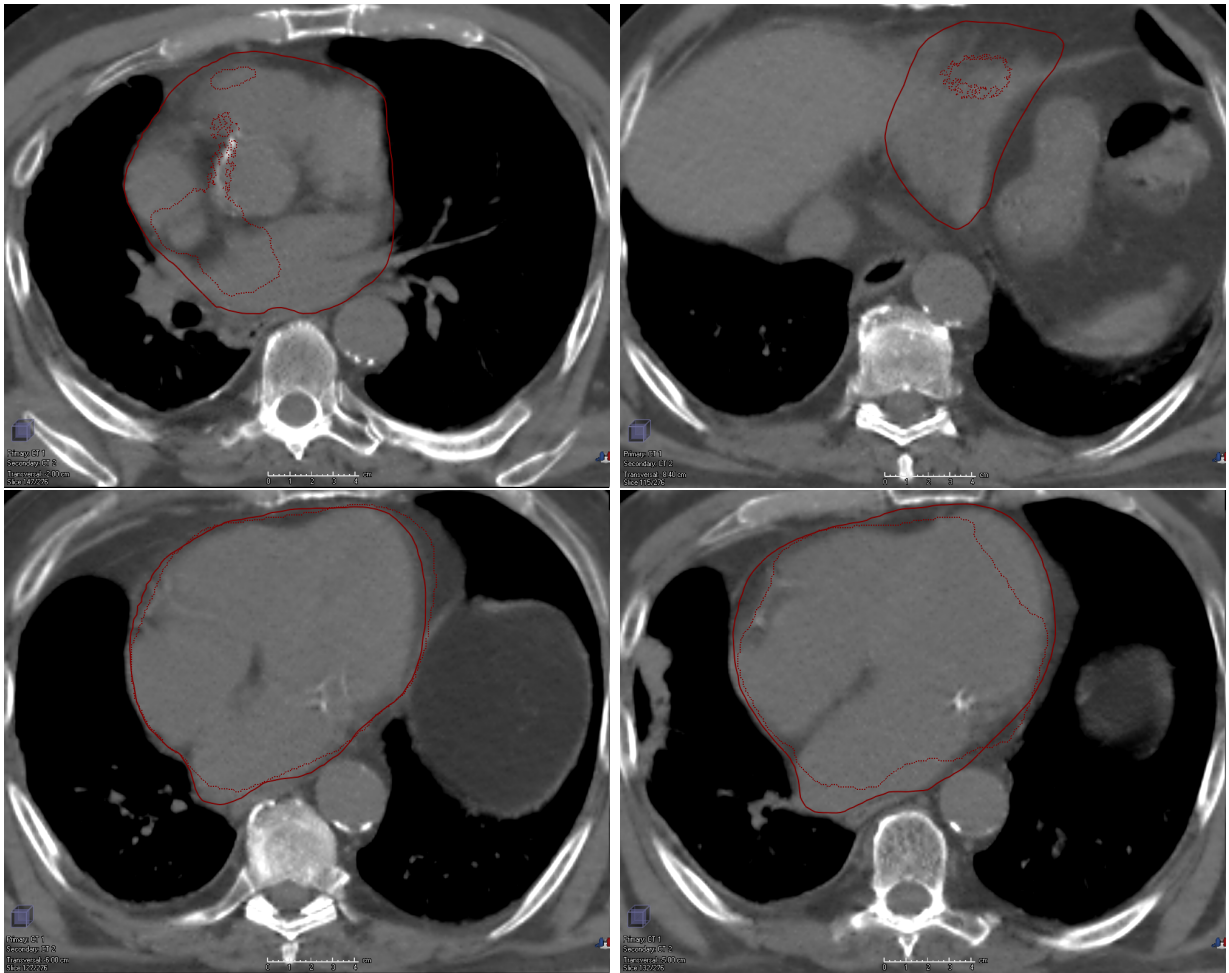
Figuur 112: Illustratie waar het model longen had geannoteerd in de darmen (testpatiënt twee) (linksboven, rechtsboven en linksonder). Willekeurig beeld waar de intekening in orde was (rechtsonder). Volle lijn: originele structuren, stippellijn: teststructuren.



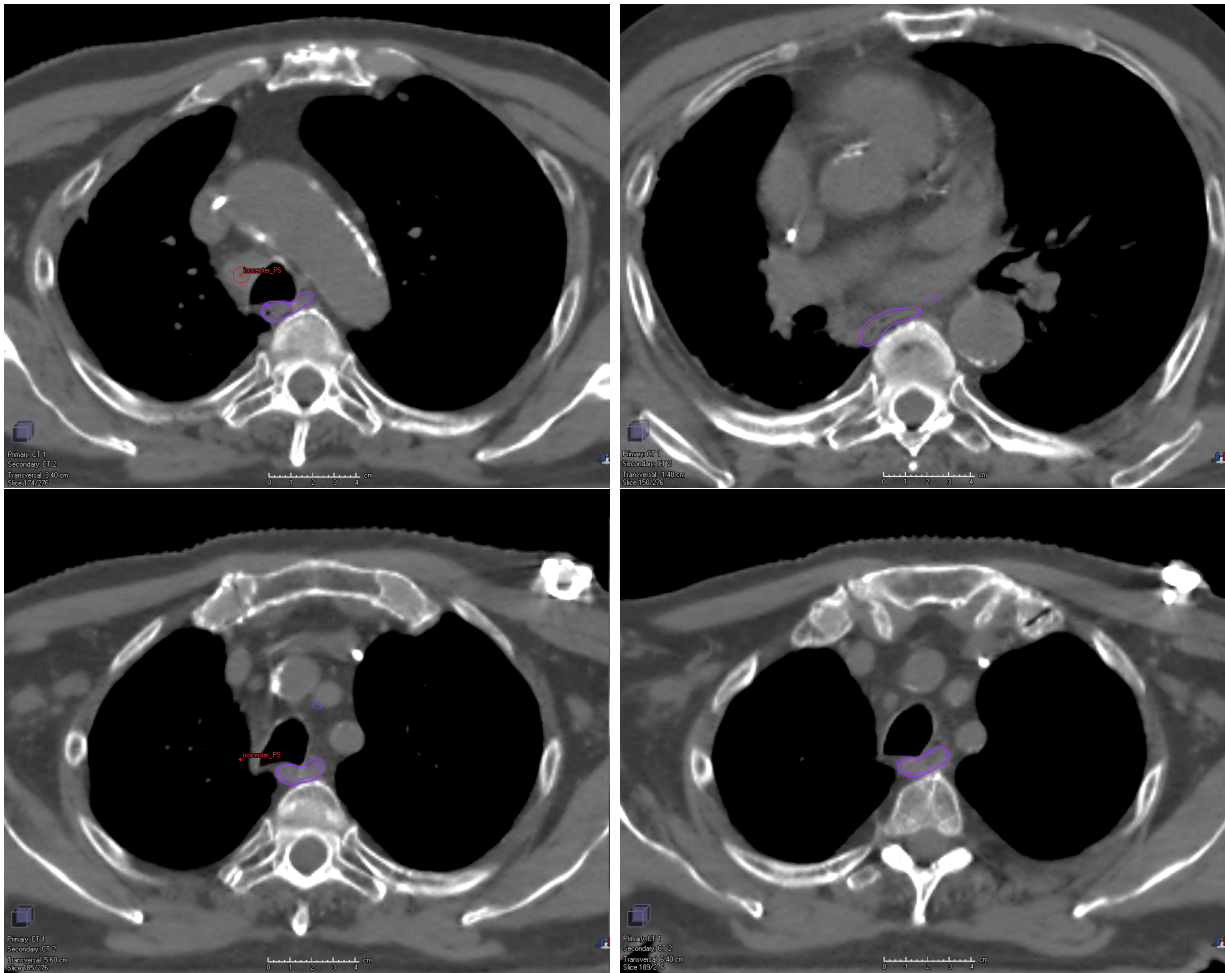
Figuur 113: Illustratie waar het model het ruggenmerg (testpatiënt twee) intekende waar er origineel geen intekening was. Hoogst ingetekende ruggenmerg (linksboven) en laagst ingetekende ruggenmerg (rechtsboven). Willekeurige beelden waar de intekeningen in orde waren (onder). Volle lijn: originele structuren, stippellijn: teststructuren.



Figuur 114: Illustratie waar het model de luchtpijp (testpatiënt twee) intekende waar er origineel geen intekening was. Hoogst ingetekende luchtpijp (linksboven) en laagst ingetekende luchtpijp (rechtsboven). Willekeurige beelden waar de intekeningen in orde waren (onder). Volle lijn: originele structuren, stippellijn: teststructuren.

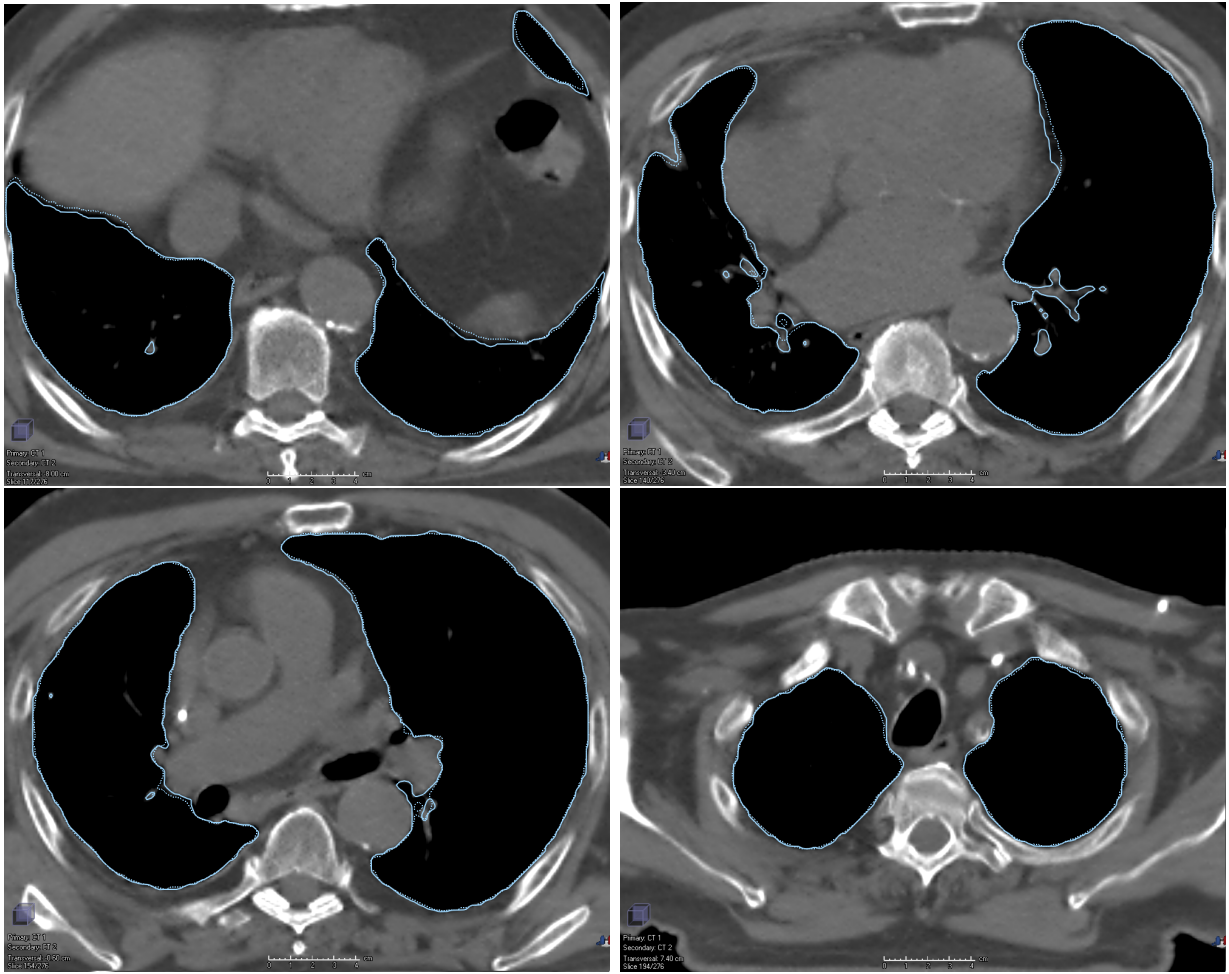


Figuur 115: Illustratie waar het intekenen niet in orde was voor het hart (testpatiënt drie) bij het begin van het orgaan (linksboven) en bij het einde van het orgaan (rechtsboven). Willekeurige beelden waar de intekening in orde was (onder). Volle lijn: originele structuren, stippellijn: teststructuren.

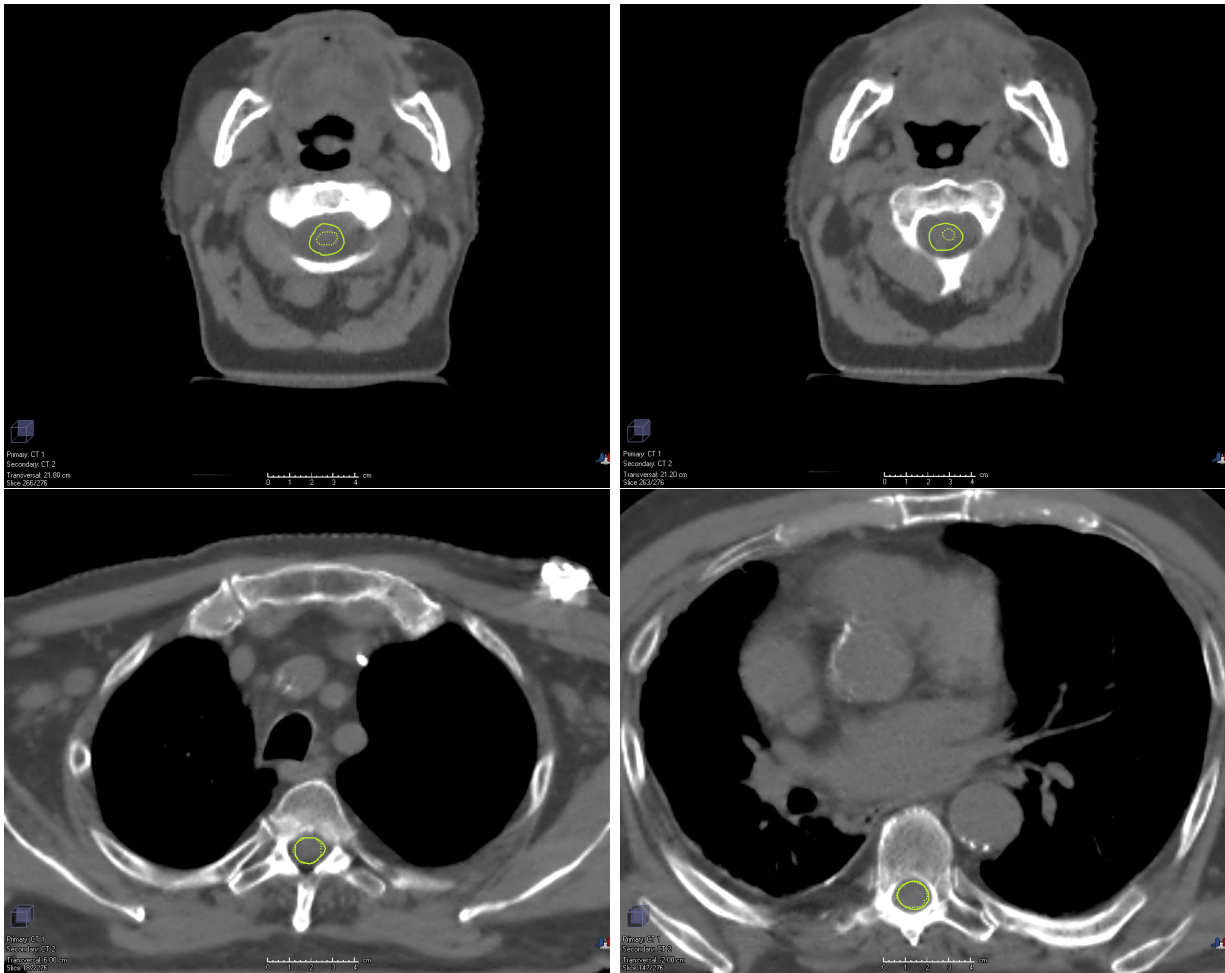


Figuur 116: Illustratie waar het intekenen niet in orde was voor de slokdarm (testpatiënt drie) in serie één (linksboven) en serie twee (rechtsboven). Beeld waar het model een extra stuk slokdarm intekende (linksonder). Willekeurig beeld waar de intekening in orde was (rechtsonder). Volle lijn: originele structuren, stippellijn: teststructuren.

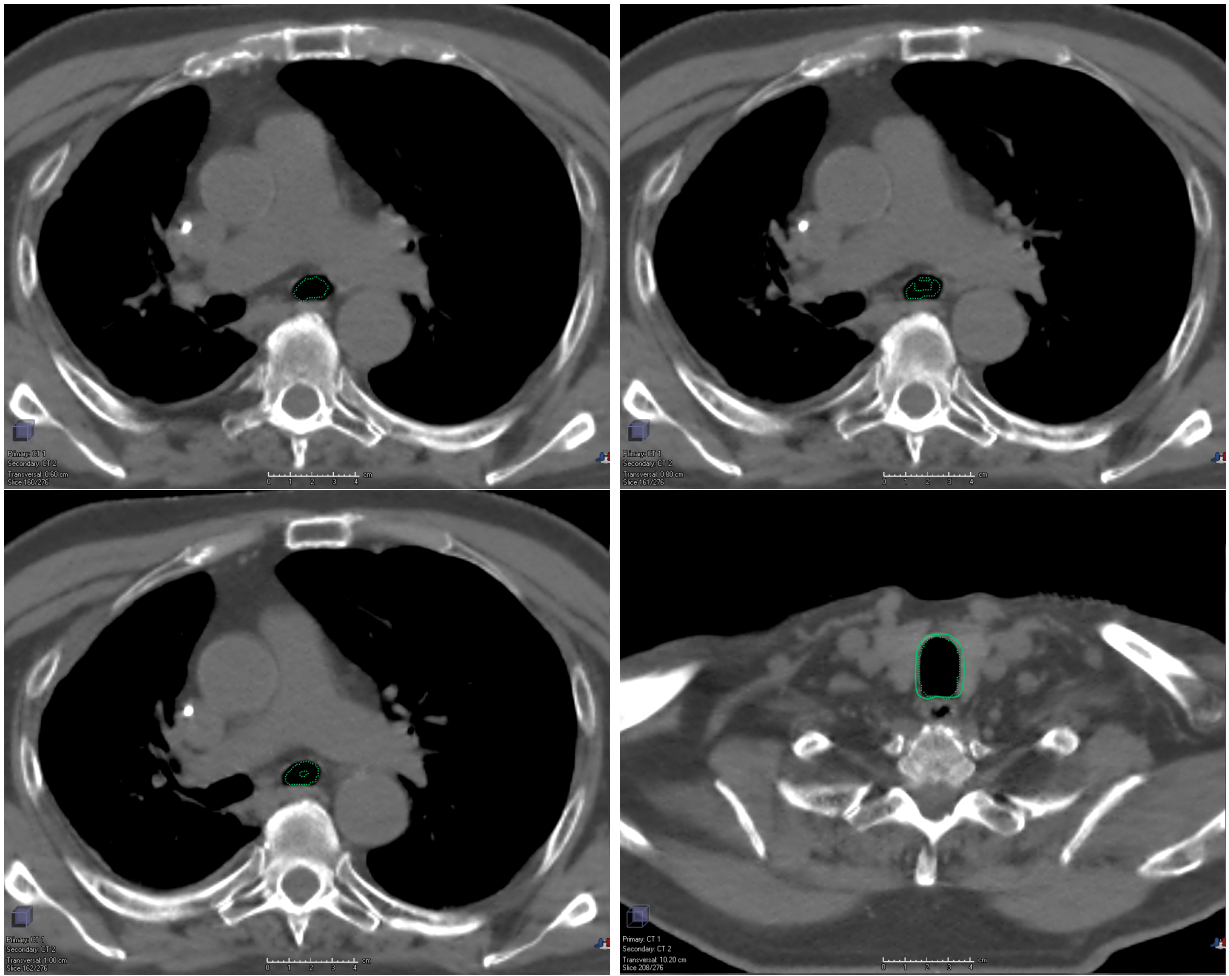




Figuur 117: Illustratie waar de intekeningen in orde waren voor de longen (testpatiënt drie). Volle lijn: originele structuren, stippellijn: teststructuren.

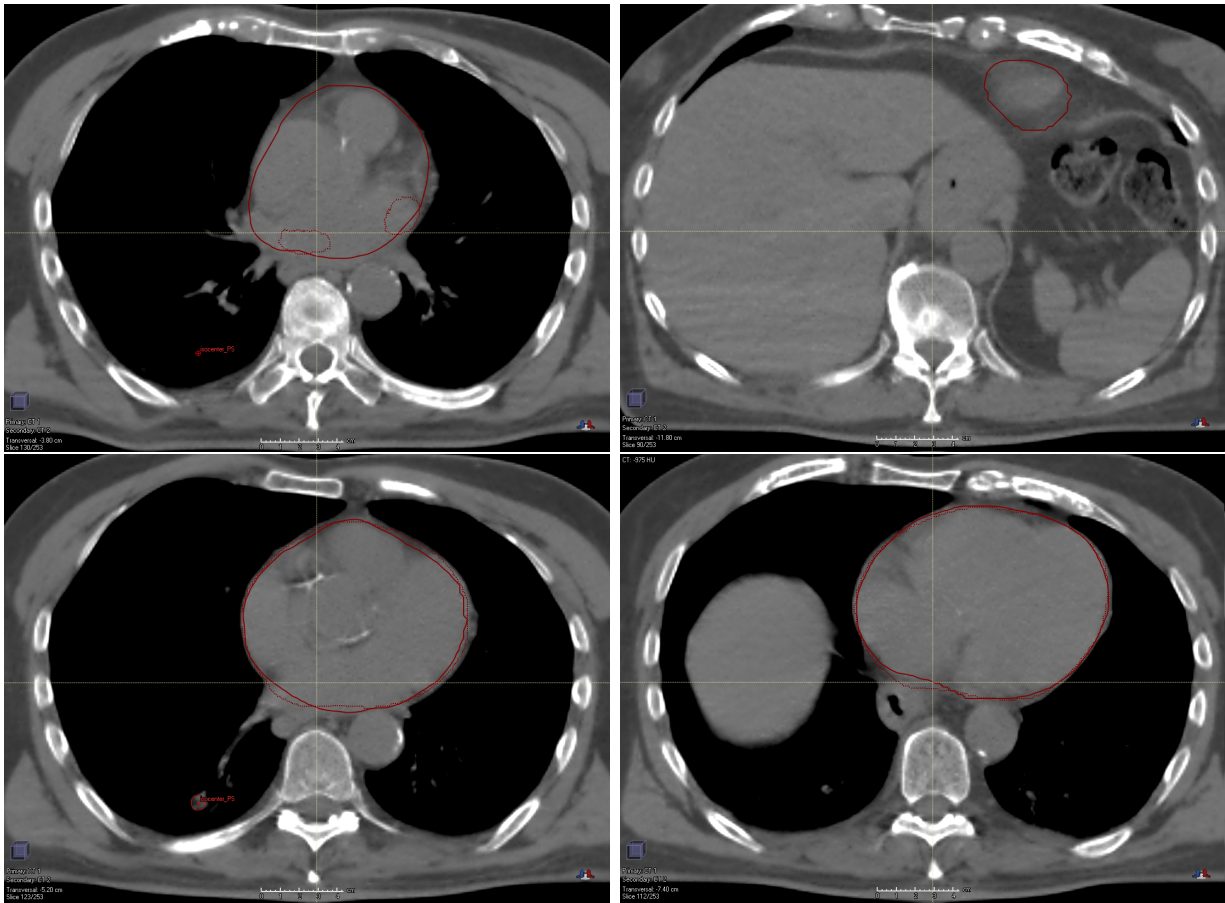


Figuur 118: Illustraties waar het model het ruggenmerg (testpatiënt drie) te klein intekende (boven). Willekeurige beelden waar de intekeningen in orde waren (onder). Volle lijn: originele structuren, stippellijn: teststructuren.

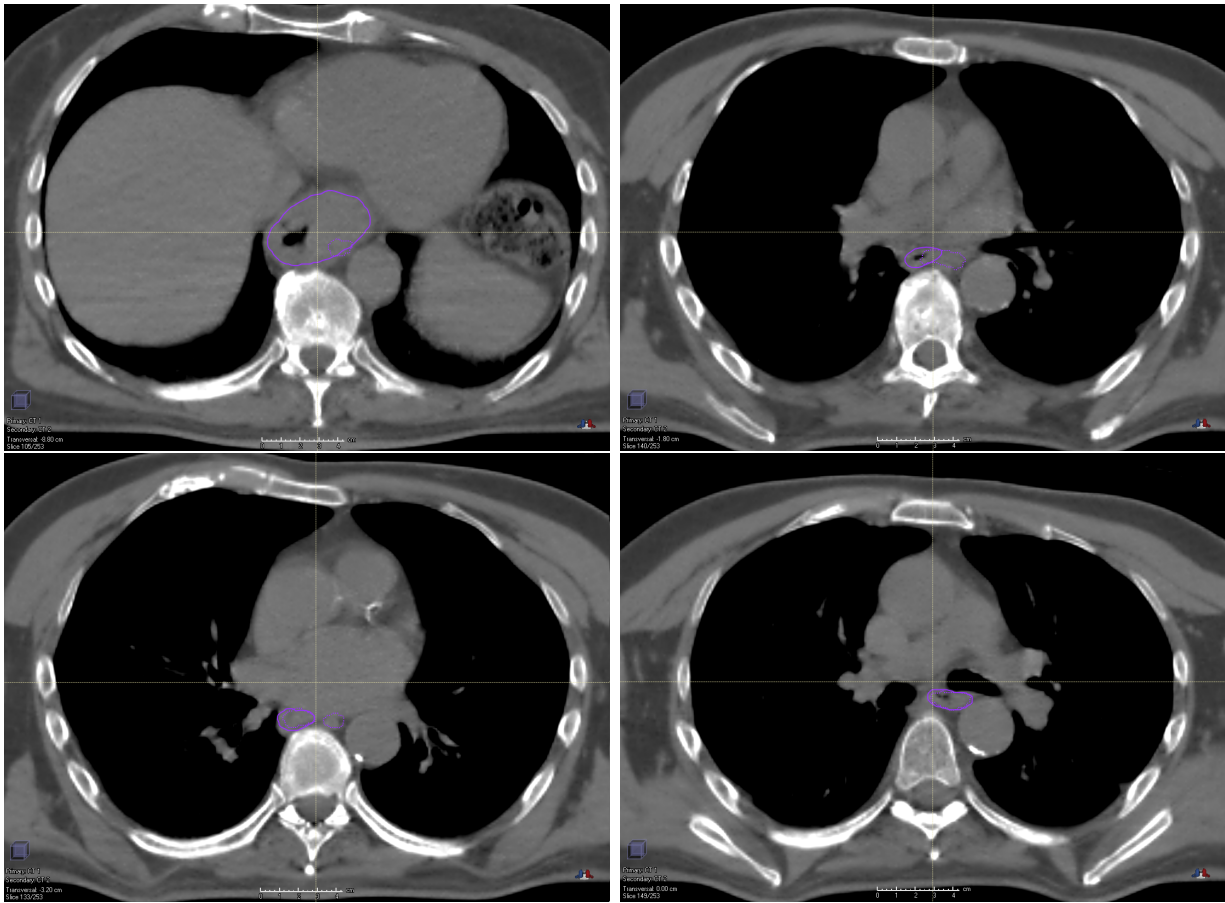


Figuur 119: Illustratie waar het model de luchtpijp (testpatiënt drie) intekende bij de linker bronchus (linksoven, rechtsboven en linksonder). Willekeurig beeld waar de intekeningen in orde waren (rechtsonder). Volle lijn: originele structuren, stippellijn: teststructuren.

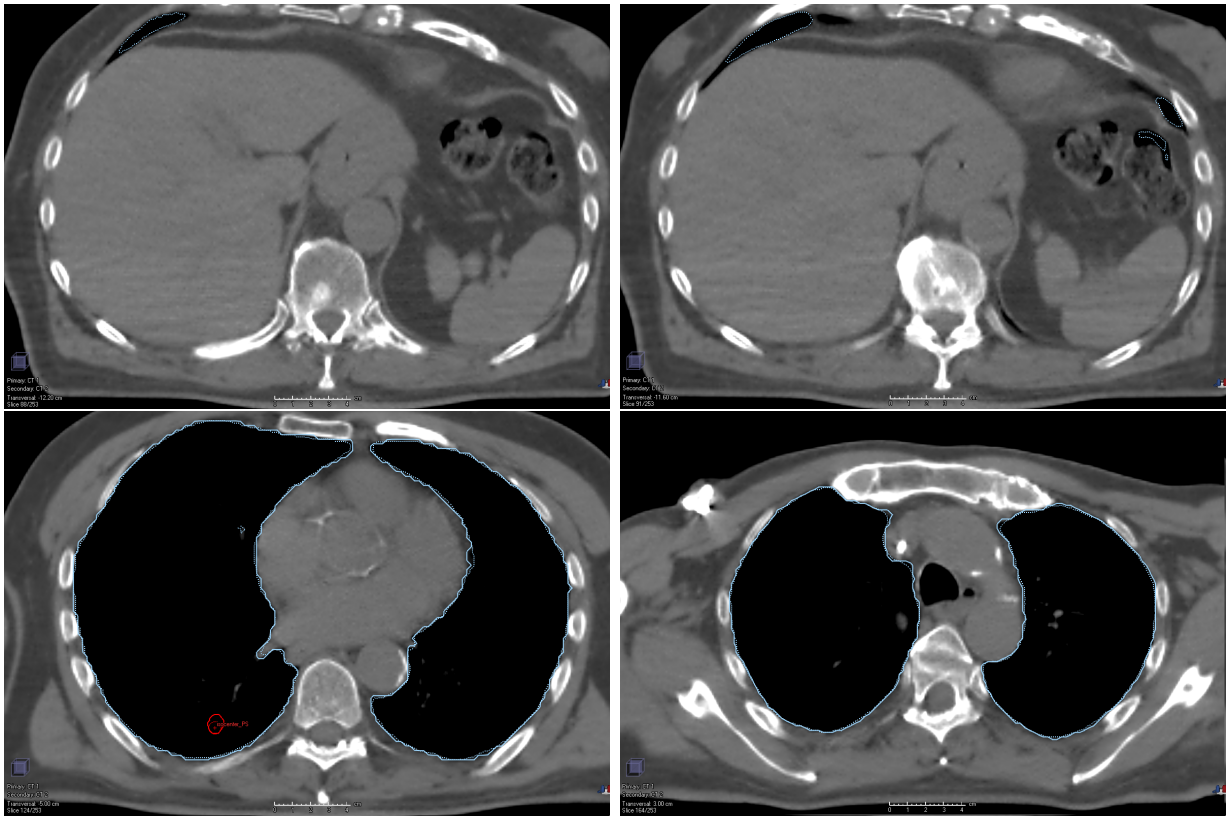




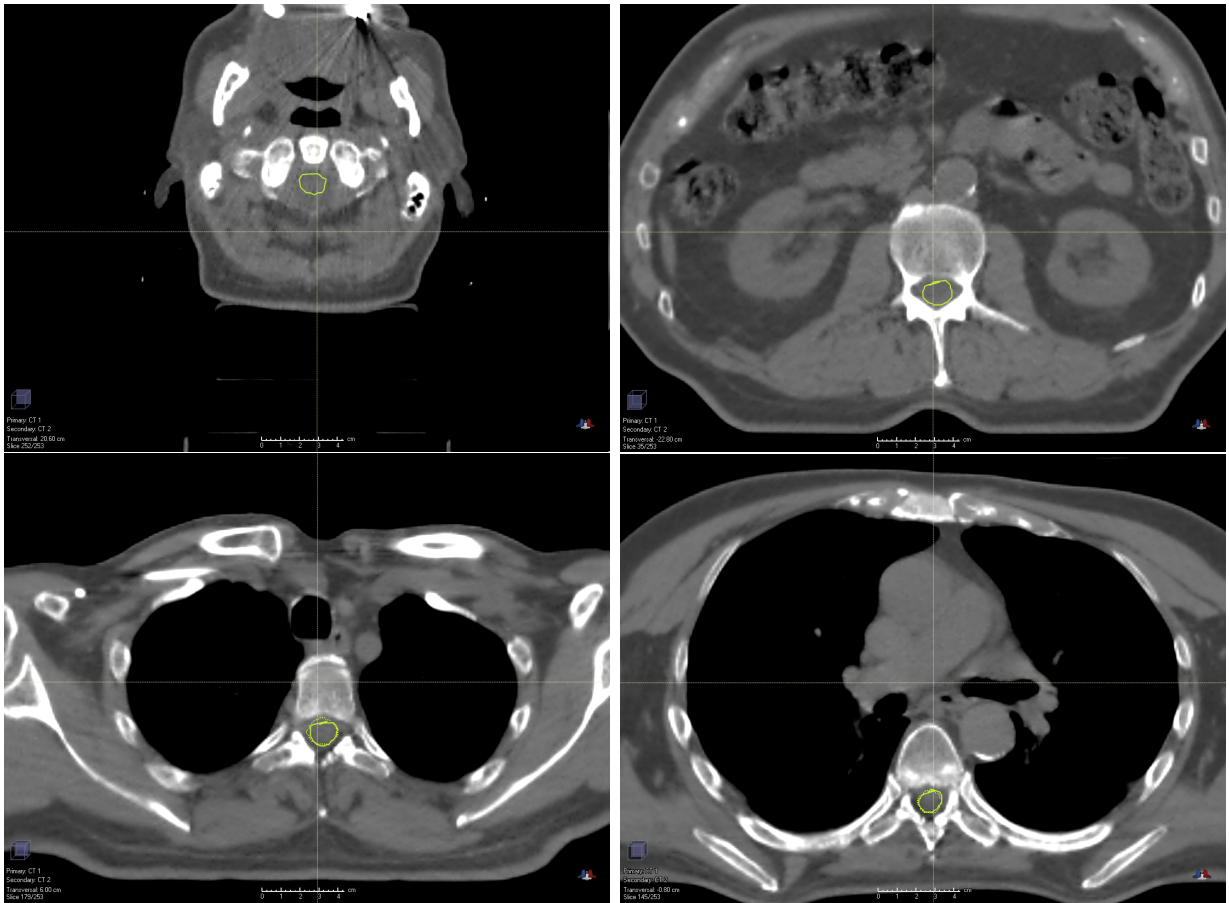
Figuur 120: Illustratie waar het intekenen niet in orde was voor het hart (testpatiënt vier) bij het begin van het orgaan (linksboven) en bij het einde van het orgaan (rechtsboven). Willekeurige beelden waar de intekening in orde was (onder). Volle lijn: originele structuren, stippellijn: teststructuren.



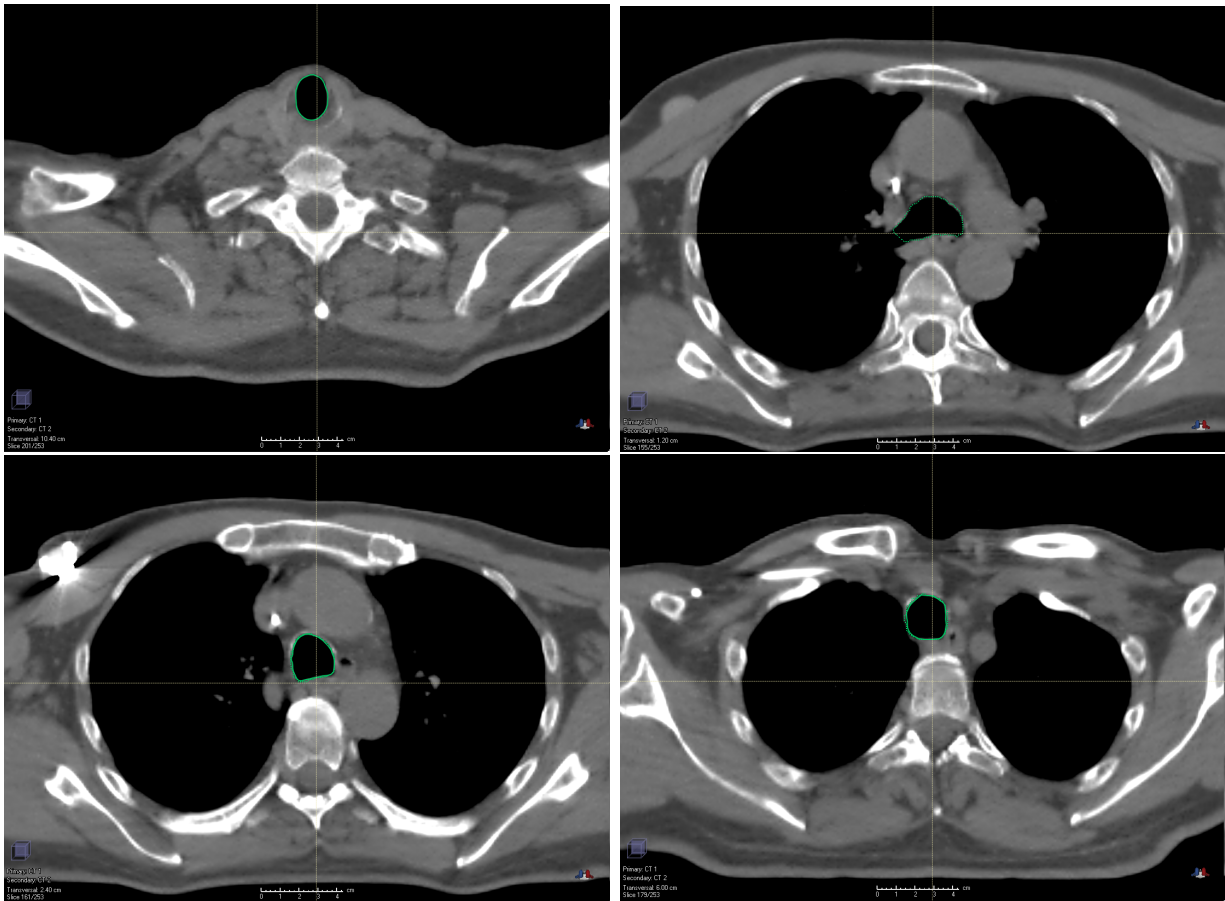
Figuur 121: Illustratie waar het intekenen niet in orde was voor de slokdarm (testpatiënt vier) in serie één (linksboven) en serie twee (rechtsboven). Beeld waar het model een extra stuk slokdarm intekende (linksonder). Willekeurig beeld waar de intekening in orde was (rechtsonder). Volle lijn: originele structuren, stippellijn: teststructuren.



Figuur 122: Illustratie waar het model nog longen voorspelde (testpatiënt vier) (boven). Wilkekeurige beelden waar de intekeningen in orde waren (onder). Volle lijn: originele structuren, stippellijn: teststructuren.



Figuur 123: Illustraties waar het model het ruggenmerg (testpatiënt vier) niet intekende in het begin van het orgaan (linksboven) en op het einde van het orgaan (rechtsboven). Willekeurige beelden waar de intekeningen in orde waren (onder). Volle lijn: originele structuren, stippellijn: teststructuren.



Figuur 124: Illustratie waar het model de luchtpijp (testpatiënt vier) niet intekende in het begin (linksboven) en te veel intekende op het einde (rechtsonder). Willekeurige beelden waar de intekeningen in orde waren (onder). Volle lijn: originele structuren, stippellijn: teststructuren.