2024-05-20

Upphandling och utvärdering av AI-baserade produkter

Angelica Svalkvist

Sjukhusfysiker, Docent

Sahlgrenska Universitetssjukhuset

VÄSTRA GÖTALANDSREGIONEN
SAHLGRENSKA UNIVERSITETSSJUKHUSET

Physica Medica 83 (2021) 257–263

Contents lists available at ScienceDirect

Physica Medica

journal homepage: www.elsevier.com/locate/ejmp

Original paper

Procurement, commissioning and QA of AI based solutions: An MPE's perspective on introducing AI in clinical practice

Hilde Bosmans [a, 1, *], Federica Zanca [b, 1], Frederik Gelaude [a]

[a] University Hospitals of the KU Leuven, Leuven, Belgium
[b] Palindromo Consulting, Leuven, Belgium

Table 1
Summary of the chronological steps in the adoption of an AI tool in healthcare, .

Terms	Scope
Procurement	To guide the selection of the optimal AI application in terms of safety, performance, match with the target use case, usability, ethical aspects and price.
Acceptance testing	To ensure compliance of a new AI application with its safety and performance specification at installation.
Commissioning	To prepare the AI application for clinical use and the roll-out within the local clinical workflow.
Quality assurance	To assure that the AI application operates over time as expected, for its purpose.

Upphandling av AI-baserad mjukvara

- Medicinsk utrustning som måste följa MDR och GDPR
- Liknande förfarande som vid upphandling av annan programvara
  - Upphandlingsgrupp - olika discipliner och kompetenser
- Säkerställa att mjukvaran uppfyller så väl säkerhetskrav som önskad prestationsnivå (på lokal inputdata)
  - Kräver detaljerad information om framtagande av mjukvaran

Upphandling

- Checklista - Juridik vid användning av AI (ES2022-08)
  - Framtagen av eSamverkan
  - Fokuserar på rättsliga frågeställningar som särskilt behöver beaktas vid användning av AI
- Några av punkterna kan ligga till grund för frågor till kravspecifikationen

Förslag på innehåll i kravspecifikation

- Information om hur träning och utvärdering av AI-algoritmen har gått till
  - Tillräcklig mängd data vid träning?
  - Typ av data som använts?
  - Finns risk för bias pga. att missvisande/icke-relevant data använts?
- Kan ett utfall/resultat från AI-algoritmen förklaras? (vad baseras utfallet på?)
- Hur kan algoritmens funktionalitet säkerställas och följas upp över tid?
- Har AI-algoritmen utvecklats med hänsyn till etik, objektivitet och saklighet?

Utvärdering - Tester och kontroller

- Acceptanstester (för att säkerställa att AI-applikationen uppfyller specifikationer)
- Kvalitetskontroller (för att säkerställa att AI-applikationen fungerar över tid)
  - Bör utföras vid flera tillfällen
    - Efter installation av testversion
    - Efter installation av ”riktig” version
    - Vid uppdateringar
    - Vid ändrad användning

[1]Bosmans et al, Procurement, commissioning and QA of AI based solutions: An MPE`s perspective on introducing AI in clinical practice, Phys Med 83:257-263, 2021.

Tester och kontroller

- Vad vill vi kontrollera?
  - Att utfallet från AI-applikationen är konsekvent och repeterbart
  - Hur oväntad eller ofullständig data hanteras
- Hur kan vi kontrollera detta?
  - Etablera en baseline

[1]Bosmans et al, Procurement, commissioning and QA of AI based solutions: An MPE`s perspective on introducing AI in clinical practice, Phys Med 83:257-263, 2021.

Etablera en baseline[1]

- Alternativ 1: Skapa en databas innehållande lokal data.
  - Viktigt att ha tillräckligt stort dataset
  - Bör även innehålla ovanliga fall
  - Facit måste vara känt (Gold standard)
  - Kan få problem med etik
- Alternativ 2: Skapa en databas innehållande simulerad testdata
  - Kan ge stor variation i testdata
  - Facit finns
  - Måste säkerställa klinisk realism
  - Är tidskrävande

[1]Bosmans et al, Procurement, commissioning and QA of AI based solutions: An MPE`s perspective on introducing AI in clinical practice, Phys Med 83:257-263, 2021.

Hur kan vi utvärdera?

- AI-baserad segmenteringsmjukvaror
  - Jämföra utfall från AI-applikation med Gold standard
    - Gold standard (facit): Manuellt ritade ROI:er
- AI-baserade beslutsstödsmjukvaror (detektion)
  - Jämföra utfall från AI-applikation med Gold standard
    - Gold standard (facit): Patologi detekterade av radiologer
- AI-baserade bildrekonstruktionsmjukvaror
  - Jämföra utfall från AI-applikation med Gold standard
    - Vad är Gold standard (facit)? Hur jämför vi?

Exempel på utvärdering av AI-baserad rekonstruktionsmjukvara

True Fidelity

- Bildrekonstruktion för DT utvecklad av GE Healthcare
- DLIR – Deep Learning Image Reconstruction
- Tränad på ett stort antal undersökningspar
  - Högdosundersökningar och lågdosundersökningar av samma objekt
  - Högdosundersökningar rekonstruerade med FBP = ”facit”
- Tre ”styrkor”: DLIR-L, DLIR-M och DLIR-H
  - Olika nivåer av brusreduktion

Utvärdering på två sätt

- VGC (visual grading characteristics)
  - Bedömning av synbarhet av anatomiska strukturer
  - 20-25 patienter
  - 3-5 granskare
  - Olika rekonstruktionsinställningar
- Mätning av fysikaliska bildkvalitetsparametrar i fantom
  - Linjäritet av CT-tal
  - Spatiell upplösning
  - Lågkontrastupplösning
  - NPS (noise power spectrum)
- Mätning av fysikaliska bildkvalitetsparameterar i patientbilder

Examensarbete Tuva Skarp[1] – VGC-studier

Bukundersökningar

- Bildkvalitetskriterier:

1. Visual reproduction of the liver parenchyma and intrahepatic vessels
2. Visual reproduction of the differentiation of the right adrenal gland from adjacent structures
3. Visual reproduction of the perirenal fat
4. Visual reproduction of the terminal ileum
5. Overall quality of the examination

	Smaller patient BMI < 25	Larger patient BMI ≥ 25
Number of patients	11	9
Male/female	7/4	0/9
Age	60	68
$CTDI_{vol}$ (mGy)	6.3 (4.9-8.2)	9.9 (5.2-21.7)
Weight (kg)	66 (50-80)	76 (57-110)
Length (cm)	171 (146-182)	158 (152-166)
BMI	22 (20-24)	30 (25-44)
AP (mm)	232 (213-251)	298 (191-387)
Lat (mm)	343 (300-376)	388 (295-505)

Hjärnundersökningar

- Bildkvalitetskriterier:

1. Visual reproduction of the border between white and grey matter
2. Visual reproduction of the basal ganglia
3. Visual reproduction of the cerebrospinal fluid space around the mesencephalon
4. Visual reproduction of the posterior fossa structures
5. Overall quality of the examination

	Born before 1940	Born after 1970	Born between 1940 and 1970
Number of patients	9	7	4
Male/female	3/6	5/2	1/3
Age	89 (82-95)	29 (16-42)	63 (52-74)
$CTDI_{vol}$ (mGy)	34.4 (29.7-41.3)	36.7 (31.0-40.2)	38.3 (34.2-41.3)

[1]Skarp, T. Evaluation of deep learning image reconstruction for brain- and abdominal CT - A visual grading characteristics study, M. Sc. Thesis. University of Gothenburg, 2021.

Resultat VGC bukundersökningar[1]

Rekonstruktionsinställningar:

ASIR-V 40% (stnd)
DLIR-M + E1 (stnd)
DLIR-H + E1 (stnd)
} 3 mm & 0.625 mm

E1 = Kantförstärkande filter

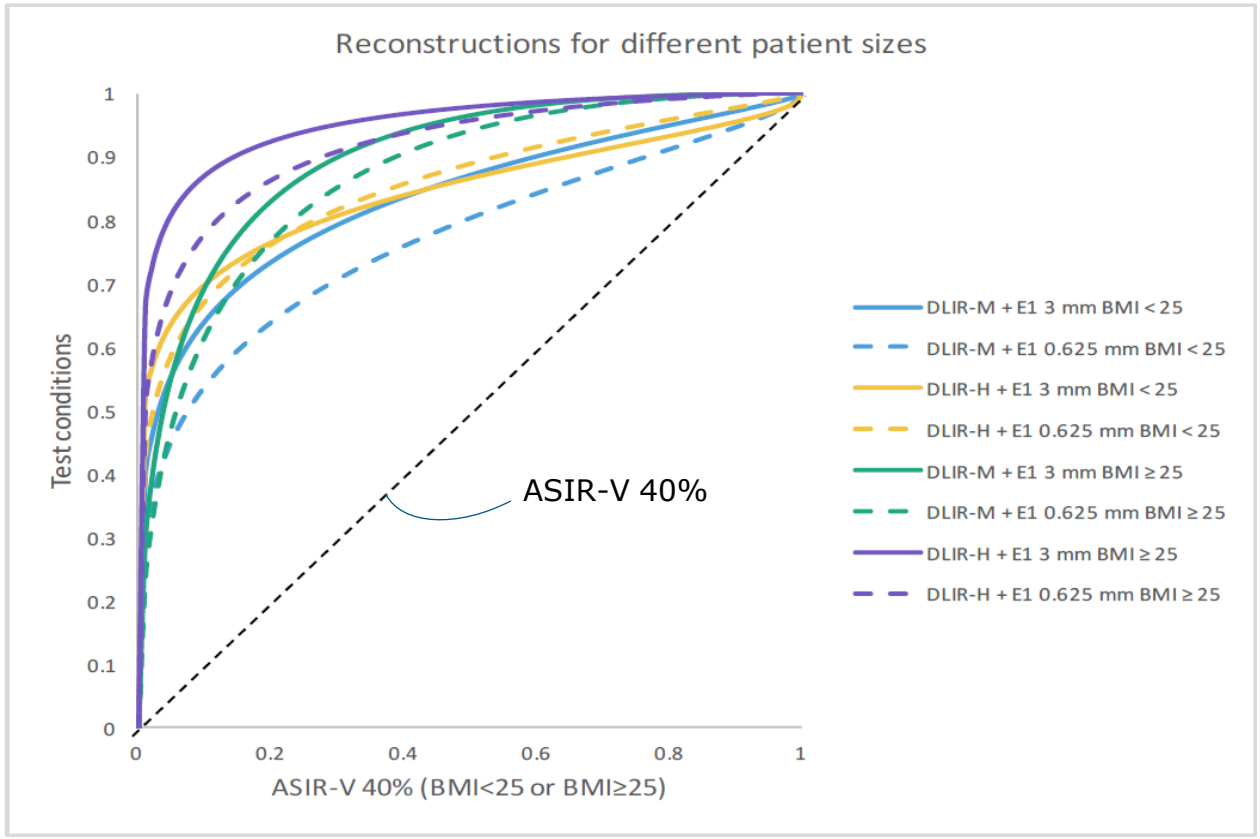


Figure 13: Pooled data for the two patient groups and their reconstructions where the solid lines show the reconstructions made with 3 mm slices filter and the dashed lines show the reconstructions made with 0.625 mm slices.

[1]Skarp, T. Evaluation of deep learning image reconstruction for brain- and abdominal CT - A visual grading characteristics study, M. Sc. Thesis. University of Gothenburg, 2021.

Resultat VGC hjärnundersökningar[1]

Rekonstruktionsinställningar:

ASIR-V 50% (soft)
DLIR-M + EC1 (stnd)
DLIR-M + EC2 (stnd)
DLIR-H + EC1 (stnd)
DLIR-H + EC2 (stnd)

} 3 mm

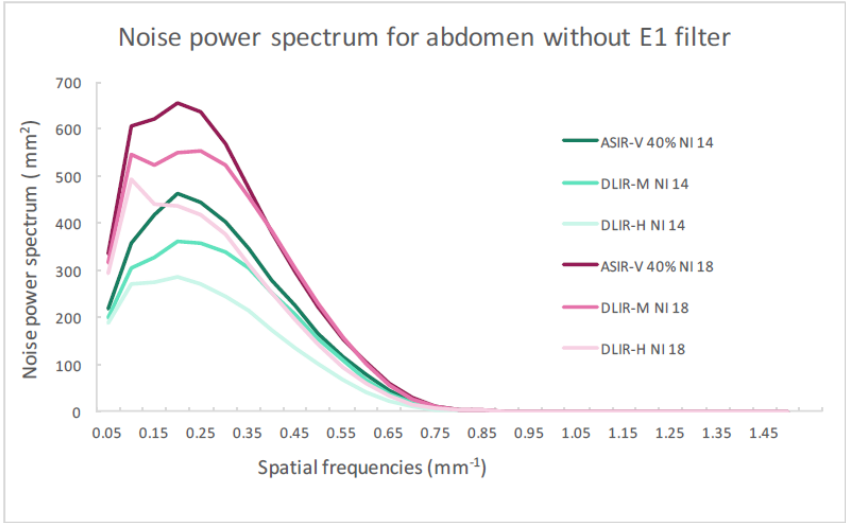
EC1 & EC2 = Kontrastförstärkande filter

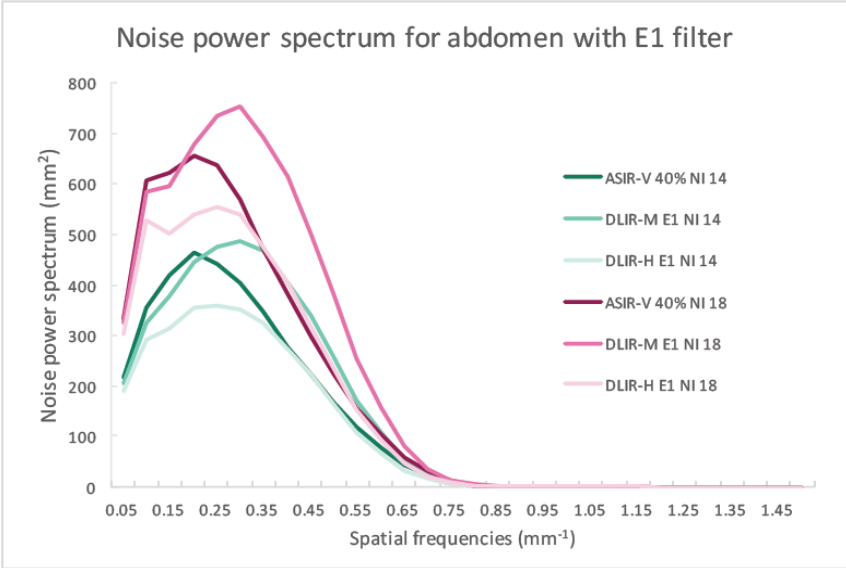
Figure 16: Pooled data for the two patient groups and their reconstructions where the solid lines show the reconstructions made with EC1 filter and the dashed lines show the reconstructions made with EC2 filter.

[1]Skarp, T. Evaluation of deep learning image reconstruction for brain- and abdominal CT - A visual grading characteristics study, M. Sc. Thesis. University of Gothenburg, 2021.

Resultat fysikalisk bildkvalitet – fantommätningar[1]

- Ingen statistiskt signifikant skillnad i HU-värde
- Ingen skillnad i spatiell upplösning
- Något högre lågkontrastupplösning i bilder rekonstruerade med DLIR
- Det kantförstärkande filtret (E1) som användes för bukundersökningarna verkar förstärka brusinnehållet i bilderna (NPS från vattenfantom)





[1]Skarp, T. Evaluation of deep learning image reconstruction for brain- and abdominal CT - A visual grading characteristics study, M. Sc. Thesis. University of Gothenburg, 2021.

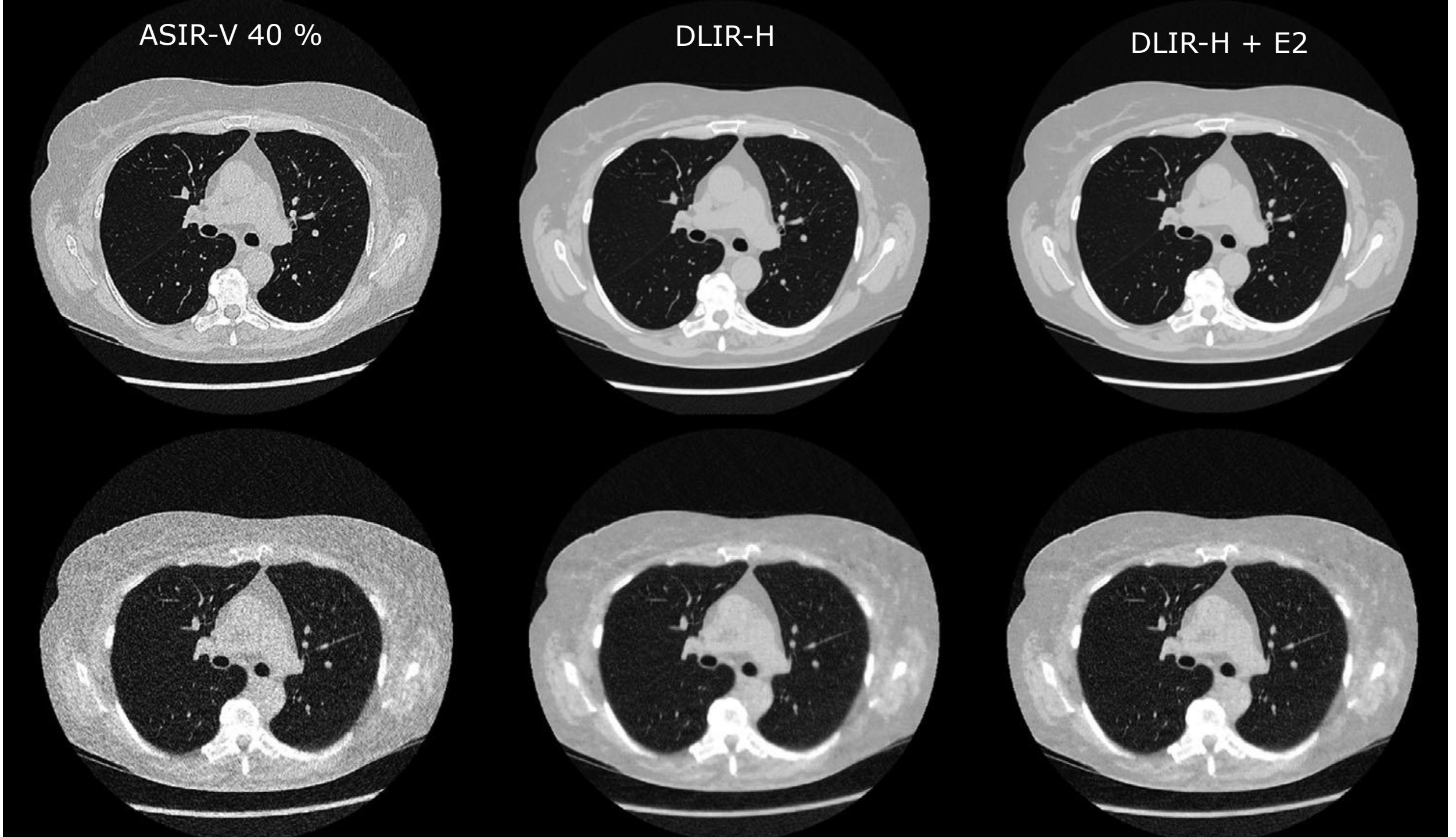
Klinisk studie – thoraxundersökningar[2]

Image quality questions		
Reproduction of anatomical structures		**Answer alternatives**
Q1.	Clear reproduction of the major fissure of the left lung (right if left one is not visible)	1. Confident that the criterion is fulfilled 2. Somewhat confident that the criterion is fulfilled 3. I do not know if the criterion is fulfilled or not 4. Somewhat confident that the criterion is not fulfilled 5. Confident that the criterion is not fulfilled
Q2.	Clear reproduction of B1: 3 subdivisions on axial plane of the apical bronchus of the right upper lobe (left upper lobe if right one is not visible)	
Q3.	Clear reproduction of A6: 4 divisions on axial plane of right apical pulmonary artery of the right lower lobe (left lower lobe if right one is not visible)	
Q4.	Clear reproduction of B6: 3 divisions on axial plane of the apical bronchus of the right lower lobe (left lower lobe if right one is not visible)	
Q5.	Clear reproduction of the right inferior pulmonary vein (RIPV): 3 divisions on axial plane (left if right one is not visible)	
Q6.	Clear reproduction of lymph node 4R (lymph node between VCS and carina)	
	General image quality	**Answer alternatives**
Q7.	Image quality acceptable for diagnosis of pulmonary nodules?	A. Fully acceptable B. Probably acceptable C. Unacceptable
Q8.	Image quality acceptable for diagnosis of fibrosis?	
Q9.	Image quality acceptable for diagnosis of emphysema?	
Q10.	Image quality acceptable for diagnosis of mediastinal inflammation (fat stranding)?	

25 patienter

Patient demographics	Mean (range)
Age (years)	66 (44-84)
Weight (kg)	77 (46-115)
Height (cm)	170 (151-188)
BMI (kg/m^2)	26 (16-34)
Radiation dose full-dose protocol	**Mean (range)**
CTDIvol (mGy)	4.5 (2.9-11.2)
DLP (mGycm)	169 (110-434)
Effective dose (mSv)	2.5 (1.6-6.3)
Radiation dose ULD protocol	**Mean (range)**
CTDIvol (mGy)	0.08 (0.07-0.1)
DLP (mGycm)	3.2 (2.7-4.2)
Effective dose (mSv)	0.05 (0.04-0.06)

[2]Svalkvist, A. et al. Evaluation of deep-learning image reconstruction for chest CT examinations at two different dose levels. J Appl Clin Med Phys, 2022;e13871.

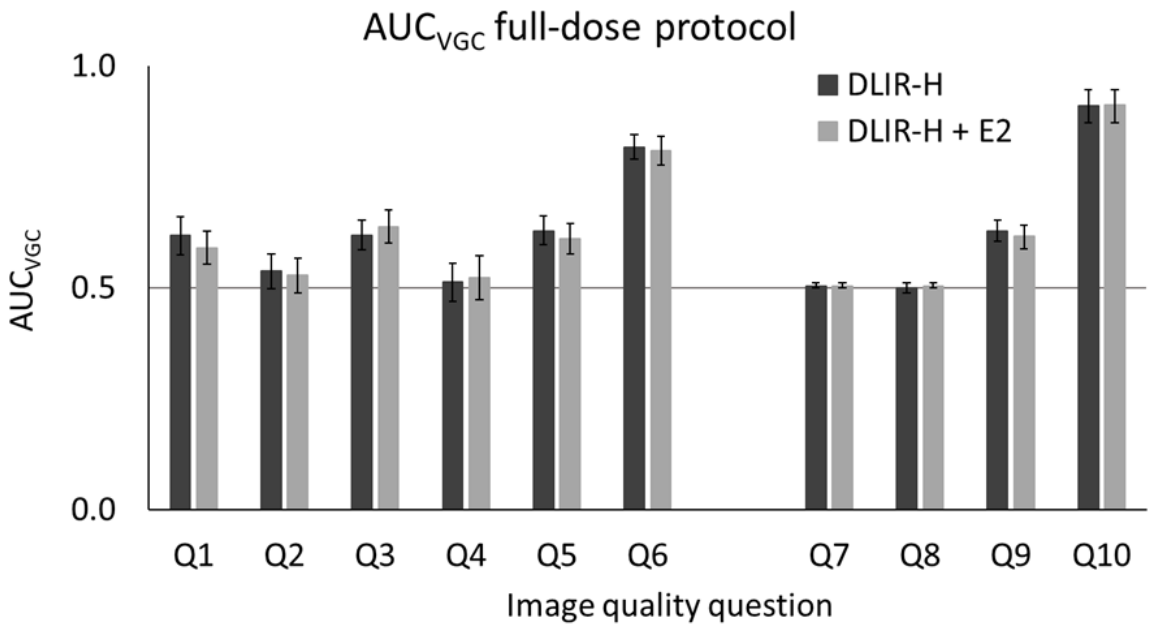

ASIR-V 40 %
DLIR-H
DLIR-H + E2

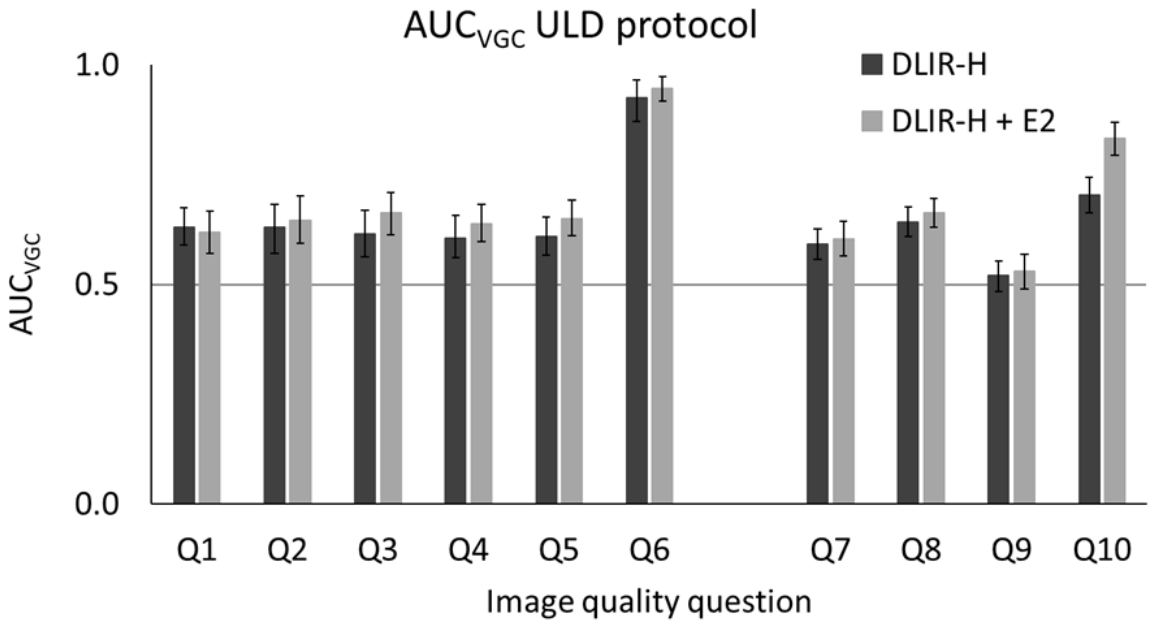
Resultat VGC-studie[2]

Rekonstruktionsinställningar:

ASIR-V 40% (lung)
DLIR-H (stnd)
DLIR-H + E2 (stnd)

} Snittjocklek 0,625 mm





[2]Svalkvist, A. et al. Evaluation of deep-learning image reconstruction for chest CT examinations at two different dose levels. J Appl Clin Med Phys, 2022;e13871.

Resultat fysikalisk bildkvalitet – kliniska bilder[2]

TABLE 4 The mean relative difference in contrast-to-noise ratio (CNR) and the standard error of the mean (SEM) for the different reconstruction settings

	Full-dose protocol		
	Mean relative difference in CNR	**SEM**	***p*-value**
DLIR-H vs. ASIR-V 40%	7.5	0.4	<0.05
DLIR-H + E2 vs. ASIR-V 40%	5.8	0.3	<0.05
DLIR-H vs. DLIR-H + E2	1.7	0.1	<0.05
	ULD protocol		
	Mean relative difference in CNR	**SEM**	***p*-value**
DLIR-H vs. ASIR-V 40%	3.0	0.1	<0.05
DLIR-H + E2 vs. ASIR-V 40%	2.3	0.1	<0.05
DLIR-H vs. DLIR-H + E2	0.6	0.04	<0.05

Note: A *p*-value < 0.05 indicates statistically significant difference.

Högre upplösning i ASIR-V 40 % beror troligtvis på att lungkernel används

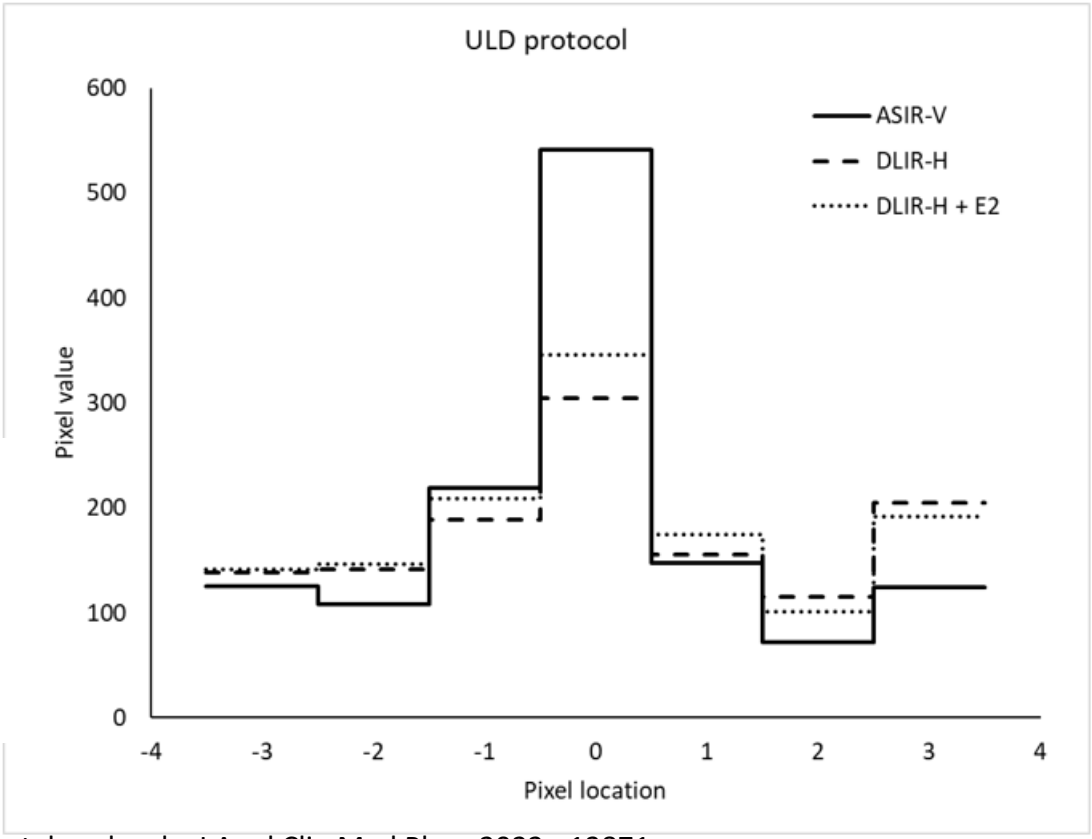
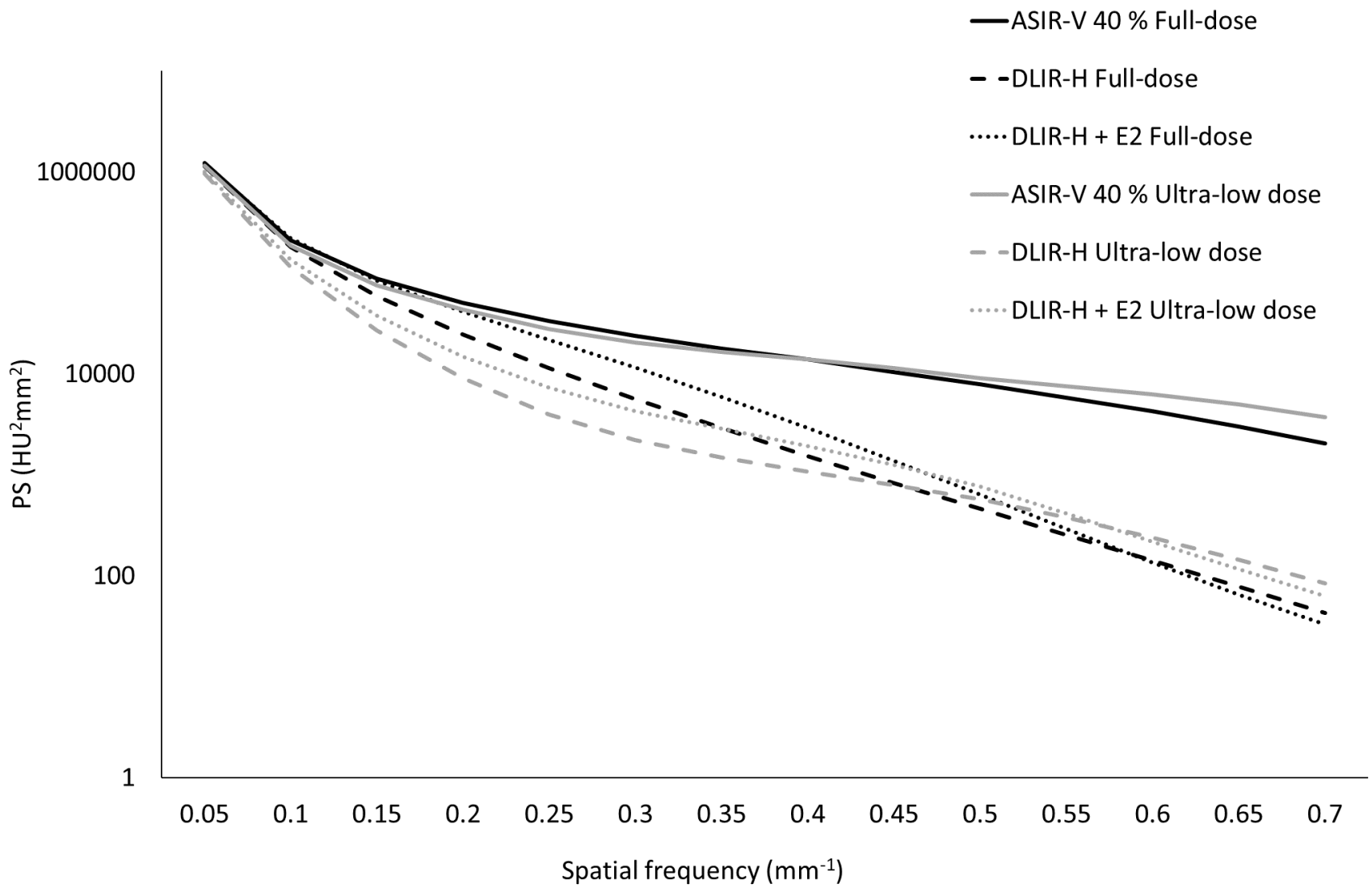


FIGURE 9 The line profiles perpendicular over a small vessel present in image slices collected using the ULD protocol. The solid line represents the line profile in the image reconstructed using ASIR-V 40%, the dashed line represents the line profile in the image reconstructed using DLIR-H and the dotted line represents the line profile in the image reconstructed using DLIR-H + E2

[2]Svalkvist, A. et al. Evaluation of deep-learning image reconstruction for chest CT examinations at two different dose levels. J Appl Clin Med Phys, 2022;e13871.

Resultat fysikalisk bildkvalitet – kliniska bilder[2]



[2]Svalkvist, A. et al. Evaluation of deep-learning image reconstruction for chest CT examinations at two different dose levels. J Appl Clin Med Phys, 2022;e13871.

Att fundera på vid utvärdering av AI-baserad rekonstuktionsmjukvara

- Fysikaliska fantom eller kliniska bilder vid utvärderingen?
  - En kombination?
- Förutsättningar för utvärderingen
  - Vad är AI-algoritmen tränad på?
  - Stråldosnivåer?
  - Patientstorlek? (Patienternas ålder?)
- Fysikaliska bildkvalitetsmått, visual grading eller detektion av patologi?
  - Vad är viktigt?
  - Vad är möjligt?
- Granskarnas erfarenheter/fördomar?


VÄSTRA
GÖTALANDSREGIONEN
SAHLGRENSKA UNIVERSITETSSJUKHUSET