Development and Application of a Web-Based Platform for Assessment of Observer Performance in Medical Imaging

Borgbjerg, Jens

Publication date:
2022

Document Version
Publisher's PDF, also known as Version of record

DEVELOPMENT AND APPLICATION OF A WEB-BASED PLATFORM FOR ASSESSMENT OF OBSERVER PERFORMANCE IN MEDICAL IMAGING

BY
JENS BORGBJERG

DISSERTATION SUBMITTED 2022

AALBORG UNIVERSITY
DENMARK

DEVELOPMENT AND APPLICATION OF A WEB-BASED PLATFORM FOR ASSESSMENT OF OBSERVER PERFORMANCE IN MEDICAL IMAGING

BY
JENS BORGBJERG

AALBORG UNIVERSITY
DENMARK

DISSERTATION SUBMITTED 2022

Dissertation submitted: August 19th, 2022

PhD supervisor: Prof. Asbjørn Mohr Drewes, MD, PhD, DMSc
Aalborg University Hospital
Aalborg University, Denmark

Assistant PhD supervisors: Prof. Jens Brøndum Frøkjær, MD, PhD
Aalborg University Hospital
Aalborg University, Denmark

Post. Doc. Heidi Søgaard Christensen, MS, PhD
Aalborg University Hospital
Aalborg University, Denmark

PhD committee: Clinical Professor Helle Damgaard Zacho (chair)
Aalborg University, Denmark

Professor Søren Rafaelsen
University of Southern Denmark, Denmark

Senior Lecturer John D. Thompson
University of Salford, UK

PhD Series: Faculty of Medicine, Aalborg University

Department: Department of Clinical Medicine

ISSN (online): 2246-1302
ISBN (online): 978-87-7573-843-4

Published by:
Aalborg University Press
Kroghstræde 3
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

Printed in Denmark by Stibo Complete, 2022

Curriculum Vitae

Jens Borgbjerg

Education and clinical work

2021-now	Consultant Radiologist, Akershus University Hospital, Norway
2016-2020	Consultant Radiologist, Aarhus University Hospital, Denmark
2012-2016	Radiology residency, Aalborg University Hospital, Denmark
2012-2012	General practitioner, Aalborg, Denmark

2011-2012	Department of Radiology, Akershus University Hospital, Norway
2010-2011	Radiology residency, Horsens Hospital, Denmark
2010-2010	Department of Pulmonary Medicine, Aarhus University Hospital, Denmark
2009-2010	Internship, Aalborg University Hospital, Denmark
2002-2009	MD, Aarhus University, Denmark

Publications

1: Borgbjerg J, Deane SF, Christensen HS, Nielsen TK. Towards Radiation Dose Reduction in Active Surveillance of Small Renal Masses: Comment on "Ultrasound Correlates Highly With Cross Sectional Imaging for Small Renal Masses in a Contemporary Cohort". Urology. 2022 Aug 6:S0090-4295(22)00640-9. doi: 10.1016/j.urology.2022.06.041. Epub ahead of print. PMID: 35944651.

2: Borgbjerg J, Nilsen FS, Nielsen TK. Unenhanced MRI for Surveillance of Small Solid Renal Masses: Additional Evidence Is Needed. AJR Am J Roentgenol. 2021 Oct;217(4):1017. doi: 10.2214/AJR.21.26063. Epub 2021 Aug 25. PMID: 34432499.

3: Borgbjerg J. Utilizing sectioned and segmented images of Visible Human datasets, comment on "Peeled images and sectioned images from real-color volume models of foot". Surg Radiol Anat. 2021 Apr;43(4):567-568. doi: 10.1007/s00276-021-02720-x. Epub 2021 Mar 10. PMID: 33689005.

4: Vedel PF, Borgbjerg J, Nielsen TK. Gangrene of the Kidney Following Percutaneous Renal Cryoablation of a Small Tumor. J Endourol Case Rep. 2020 Dec 29;6(4):490-492. doi: 10.1089/cren.2020.0139. PMID: 33457710; PMCID: PMC7803205.

5: Borgbjerg J. Web-based imaging viewer for real-color volumetric reconstruction of human visible project and DICOM datasets. Clin Anat. 2021 Apr;34(3):470-477. doi: 10.1002/ca.23712. Epub 2021 Jan 7. PMID: 33347648.

6: Christensen HS, Borgbjerg J, Børty L, Bøgsted M. On Jones et al.'s method for extending Bland-Altman plots to limits of agreement with the mean for multiple observers. BMC Med Res Methodol. 2020 Dec 11;20(1):304. doi: 10.1186/s12874-020-01182-w. PMID: 33308154; PMCID: PMC7730774.

7: Nielsen TK, Vedel PF, Borgbjerg J, Andersen G, Borre M. Renal cryoablation: five- and 10-year survival outcomes in patients with biopsy-proven renal cell carcinoma. Scand J Urol. 2020 Oct;54(5):408-412. doi: 10.1080/21681805.2020.1794954. Epub 2020 Jul 22. PMID: 32700594.

8: Borgbjerg J, Bylling T, Andersen G, Thygesen J, Mikkelsen A, Nielsen TK. CT- guided

cryoablation of renal cancer: radiation burden and the associated risk of secondary cancer from procedural- and follow-up imaging. Abdom Radiol (NY). 2020 Nov;45(11):3581-3588. doi: 10.1007/s00261-020-02527-1. PMID: 32285178.

9: Borgbjerg J, Hørlyck A. Web-Based GPU-Accelerated Application for Multiplanar Reconstructions from Conventional 2D Ultrasound. Ultraschall Med. 2021 Apr;42(2):194-201. English. doi: 10.1055/a-0999-5347. Epub 2019 Sep 5. PMID: 31487752.

10: Borgbjerg J. MULRECON: A Web-based Imaging Viewer for Visualization of Volumetric Images. Curr Probl Diagn Radiol. 2019 Nov-Dec;48(6):531-534. doi: 10.1067/j.cpradiol.2018.09.001. Epub 2018 Sep 19. PMID: 30340913.

11: Borgbjerg J, Bøgsted M, Lindholt JS, Behr-Rasmussen C, Hørlyck A, Frøkjær JB. Superior Reproducibility of the Leading to Leading Edge and Inner to Inner Edge Methods in the Ultrasound Assessment of Maximum Abdominal Aortic Diameter. Eur J Vasc Endovasc Surg. 2018 Feb;55(2):206-213. doi: 10.1016/j.ejvs.2017.11.019. Epub 2017 Dec 23. PMID: 29277483.

12: Borgbjerg J, Madsen F, Odgaard A. Patient Self-Assessed Passive Range of Motion of the Knee Cannot Replace Health Professional Measurements. J Knee Surg. 2017 Oct;30(8):829-834. doi: 10.1055/s-0037-1598174. Epub 2017 Mar 1. PMID: 28249347.

List of Papers

This thesis was based on the following papers:

I. **Borgbjerg J**, Bøgsted M, Lindholt JS, Behr-Rasmussen C, Hørlyck A, Frøkjær JB. Superior Reproducibility of the Leading to Leading Edge and Inner to Inner Edge Methods in the Ultrasound Assessment of Maximum Abdominal Aortic Diameter. Eur J Vasc Endovasc Surg. 2018 Feb;55(2):206-213. doi: 10.1016/j.ejvs.2017.11.019. Epub 2017 Dec 23. PMID: 29277483.

II. **Borgbjerg J**. MULRECON: A Web-based Imaging Viewer for Visualization of Volumetric Images. Curr Probl Diagn Radiol. 2019 Nov-Dec;48(6):531-534. doi: 10.1067/j.cpradiol.2018.09.001. Epub 2018 Sep 19. PMID: 30340913.

III. **Borgbjerg J**. Christensen HS, Al-Mashhadi R, Bøgsted M, Froekjaer J, Medrud L, Larsen N, Pedersen M, Thygesen J, Sivesgaard K, Lindholt J. Ultra-low-dose non-contrast CT is an adequate replacement for CT angiography in the assessment of maximal abdominal aortic diameter - [resubmitted (major revision) to Acta Radiologica Open]

IV. **Borgbjerg J**. Steinkohl E, Olesen S, Akisik F, Bethke A, Bieliuniene E, Christensen H, Engjom T, Haldorsen I, Kartalis N, Lisitskaya M, Naujokaite G, Novovic S, Ozola-Zālīte I, Phillips A, Swensson J, Drewes A, Frøkjær J. Observer Variability of Parenchymal- and Ductal Diameters in Chronic Pancreatitis: A Multi-institutional Study of CT Images - [resubmitted to Abdominal Radiology]

Peer-reviewed publications related to but not part of the thesis:

- Christensen HS, **Borgbjerg J**, Børty L, Bøgsted M. On Jones et al.'s method for extending Bland-Altman plots to limits of agreement with the mean for multiple observers. BMC Med Res Methodol. 2020 Dec 11;20(1):304. doi: 10.1186/s12874-020-01182-w. PMID: 33308154; PMCID: PMC7730774.

- **Borgbjerg J**, Hørlyck A. Web-Based GPU-Accelerated Application for Multiplanar Reconstructions from Conventional 2D Ultrasound. Ultraschall Med.

2021 Apr;42(2):194-201. English. doi: 10.1055/a-0999-5347. Epub 2019 Sep 5. PMID: 31487752.

- **Borgbjerg J**. Web-based imaging viewer for real-color volumetric reconstruction of human visible project and DICOM datasets. Clin Anat. 2021 Apr;34(3):470-477. doi: 10.1002/ca.23712. Epub 2021 Jan 7. PMID: 33347648.

Other peer-reviewed conference abstracts related to but not part of the thesis:

- **Borgbjerg, J**. Novel web-based tool for conducting observer performance studies in imaging research. (2016) (https://epos.myesr.org/poster/esr/ecr2016/C-1635).

Abbreviations

AAA	Abdominal aortic aneurysm
PDbody	Largest diameter of the pancreatic body
PDhead	Largest diameter of the pancreatic head
ALARA	As Low As Reasonably Achievable
CI	Confidence interval
CP	Chronic pancreatitis
CPU	Central processing unit
CT	Computed Tomography
DICOM	Digital Imaging and Communications in Medicine
GPU	Graphics processing unit
HTML5	Hyper Text Markup Language 5
ITI	Inner to inner edge
ICC	Intra-class correlation coefficient
JSON	JavaScript Object Notation
LTL	Leading to leading edge
LoAs	Limits of agreement
LOAM	Limits of agreement with the mean
Dmax	Maximum abdominal aortic diameter
MRI	Magnetic Resonance Imaging
MPDhead	Largest main pancreatic duct diameter in the pancreatic head
MPDbody	Largest main pancreatic duct diameter in the pancreatic body
MPR	Multiplanar reconstructions
OTO	The outer to the outer edge
PACS	Picture Archiving and Communication System
PHP	Hypertext Preprocessor
WOAP	Web-based platform for observer agreement studies
ROC	Receiver operating characteristics
WebGL	Web Graphics Library
SQL	Structured Query Language
SUS	System Usability Scale
SBPC	The Scandinavian Baltic Pancreatic Club
ULDNC-CT	Ultra-low-dose non-contrast CT

VIVA	Viborg Vascular screening trial

English Summary

In imaging research and development, it is well-recognized that assessment of observer performance has a vital role in advancing radiology practice, technique, training, and quality control. Biomarkers derived from medical imaging studies are commonly used as decision-making tools in both clinical trials and routine clinical practice. Consequently, knowing the range of observer variation is of utmost importance to gauge if a change in an imaging biomarker is real - and reflects treatment or disease progression - or if it is due to possible observer variation. In addition, assessment of agreement is vital to establish if a dose-saving imaging protocol is feasible for a given diagnostic task. Imaging agreement studies can be performed retrospectively and are, in principle, relatively easy to perform. However, the radiological community has emphasized that agreement studies are too rarely conducted. Many studies employ an insufficient scale and scope of observers to allow enough generalizability to translate results into clinical practice. This problem can largely be explained by the logistical challenges associated with conducting observer agreement studies that span a broad range of observer experience from different institutions.

The overall aim of this Ph.D. thesis was to develop and apply a web-based platform for the facilitation of imaging observer performance studies with an automatic gathering of quantitative data from case assessments. A particular focus was to evaluate the measurement variability of uni-dimensional continuous variables, explored in three application studies. In addition, the thesis encompasses one method study.

The first application study, reported in paper I, explored the principal problem of establishing superiority in terms of measurement reproducibility among different imaging methods used in clinical practice. More specifically, this study used an initial version of the web-based platform to assess reproducibility in ultrasonographic assessment of maximum abdominal aortic diameter with three methods of caliper placements. Results from the evaluation of static two-dimensional ultrasound images by 18 observers from different institutions indicate that the method where calipers are placed from the inner wall to the inner wall of the aorta is the superior method in terms of reproducibility.

The method study is reported in paper II, and it describes the development of a web-based imaging viewer to visualize volumetric datasets. In the second application study, reported in paper III, the aforementioned imaging viewer was coupled with the web-based platform to explore a case of the general problem of investigating whether one imaging modality can replace another for a given clinical task. More specifically, the interchangeability of ultra-low-dose non-contrast computed tomography (CT) and standard-dose CT

angiography in terms of accuracy and reproducibility in determining maximum abdominal aortic diameter was evaluated in a single-center study employing 7 observers with varying experience levels. The study found that non-contrast CT scans at ultra-low-dose levels are interchangeable with gold-standard CT angiography to assess abdominal aortic diameter.
The third application study, reported in paper IV, quantified the level of observer variability in CT-based measurements of ductal- and gland diameters in chronic pancreatitis. The study encompassed 16 observers from 10 different institutions as well as 10 countries. We concluded that two-point pancreatic measurements are subject to substantial intra- and interobserver variability even among specialists.
In addition, the usability of the developed platform was evaluated with the industry-standard System Usability Scale. The usability of the developed platform and imaging viewer compared favorably to well-known products with high usability. Assessment of rendering and download speed was evaluated and found sufficient for conducting observer performance studies.
Overall, the developed web-based platform proved helpful in facilitating observer performance studies in accordance with recommendations stipulated in the radiological research literature.

Dansk Resumé

Indenfor radiologisk forskning og udvikling er det velkendt at bestemmelse af observatørvariation har en vigtig rolle i forhold til klinisk praksis, uddannelse og kvalitetskontrol. Billeddiagnostiske biomarkører anvendes ofte i beslutningeprocesser både i kliniske forsøg såvel som i klinisk praksis. Derfor er det afgørende at kende spændvidden af den forventede observatør variation således man kan vurdere om en ændring i en biomarkør afspejler en reel biologisk forskel eller den blot repræsenterer målevariation. Derudover er bestemmelse af observatør variation også afgørende for hvorvidt man kan anvende f.eks en lav-dosis computed tomography (CT) protokol i stedet for en fuld dosis til besvarelse af en given klinisk problemstilling. Fordi observatør variations studier kan udføres med retrospektive data er de i princippet nemme at gennemføre. Det radiologiske forskningsmiljø har imidlertid understreget, at studier som undersøger observatør variation er sjældne. Hertil kommer at de studier der gennemføres i majoriteten af tilfælde inkluderer observatører i et antal og spænd af erfaring således at generaliserbarheden er utilstrækkelig. Dette problem kan i en vis udstrækning forklares med baggrund i de logistiske udfordringer, der er forbundet med at udføre studier, der inkluderer en bred vifte af observatørers fra forskellige institutioner.

Hovedformålet med denne ph.d. afhandling var at udvikle og anvende en webbaseret platform til facilitering af radiologiske observatør variations studier. Et særligt fokus var at bestemme måle variabiliteten i forbindelse med endimensionelle kontinuerte variabler. Dette blev undersøgt i tre anvendelsestudier. Derudover blev der gennemført et metodestudie.

Den første anvendelses studie, rapporteret i artikel I, udforskede det principielle problem med at etablere hvorvidt en målemetode indenfor billeddiagnostik har bedre reproducerbarhed end andre der anvendes i klinisk praksis. Til dette studie anvendtes en version af den webbaserede platform til at vurdere reproducerbarhed af metoder til målemarkør placering i ultralyds bestemmelsen af maksimal abdominal aorta diameter. I studiet vurderede 18 observatører statiske to-dimensionelle ultralydsbilleder af abdominal aorta og resultaterne indikerer at placering af målemarkør fra den indre væg til den indre væg af aorta har bedst reproducerbarhed.

Metodestudiet rapporteret i artikel II beskriver udviklingen af en webbaseret DICOM viewer som kan anvendes til visualisering af radiologiske volumetriske datasæt.

I det andet anvendelses studie, rapporteret i artikel III, blev den førnævnte DICOM viewer koblet sammen med den webbaserede platform med henblik på at gennemføre et studie som er eksponent for det generelle problem med at sandsynliggøre hvorvidt en

billeddiagnostisk modalitet kan udskiftes med en anden. Mere specifikt blev det undersøgt hvorvidt en standard-dosis CT angiografi kan erstattes af en ultra-lav-dosis ikke-kontrast CT til bestemmelse af maksimal abdominal aorta diameter. Dette studie inkluderede 7 observatører med forskellige erfaringsniveauer og fandt at førnævnte undersøgelsestyper er udskiftbare med hinanden i forhold til denne klinisk problemstilling.
Det tredje anvendelses studie, rapporteret i artikel IV, bestemte omfanget af observatør måle variabilitet når der udføres CT-baserede målinger af duktale- og parenkymdiametre hos kronisk pankreatit.patienter. Studiet inkluderede 16 observatører fra 10 forskellige institutioner og 10 lande. Selv blandt specialister fandt vi betydelig intra- og interobservatør variation af sådanne målinger. Brugervenligheden af den udviklede platform blev evalueret med et spørgeskema i form af industristandarden System Usability Scale. Brugervenligheden af den udviklede platform og DICOM viewer blev vurderet til at være på linje med velkendte produkter som er anerkendt for at have høj brugervenlighed. Vurdering af DICOM viewerens grafikafviklings- og downloadhastighed blev evalueret og fundet tilstrækkelig til at udføre radiologiske observatør variations studier.
Samlet set viste den udviklede webbaserede platform sig nyttig til lettere at gennemføre radiologiske observatører variations undersøgelser i overensstemmelse med anbefalingerne fremsat i den radiologiske forskningslitteratur.

Acknowledgments

I wish to express my sincere gratitude and appreciation to all those who have contributed to and made this work possible. In particular:
Professor Asbjørn Mohr Drewes, my main supervisor, for enthusiastic scientific coaching.
Professor Jens Brøndum Frøkjær, my co-supervisor, for inspiring tutoring and friendly, rewarding support.
Heidi Søgaard Christensen, my co-supervisor, and Professor Martin Bøgsted for collaboration in refining the limits of agreement with the mean statistical method and excellent statistical discussions.
Professor Jes S. Lindholt for introducing me to abdominal aortic aneurysm research through collaboration on two thesis papers.
My wonderful colleagues at the Departments of Radiology at Aarhus University Hospital, Aalborg University Hospital, and Akershus University Hospital have provided valuable formal and informal feedback during the development and application of the web-based platform.
My current employer, Akershus University Hospital, has provided me time off from clinical duties to attend Ph.D. courses and allocated time in my ongoing employment to expand upon the research at hand.
All of the 41 health professionals who have invested time and effort in participating as observers in the application studies of this Ph.D.
My family, for their support and for believing in me.

Jens Borgbjerg
Oslo, May 16, 2022

Table of Contents

Table of Contents 1

List of Figures 3

Introduction 5
- Medical imaging efficacy 5
- Observer performance studies in imaging research 7
- The volumetric imaging revolution 9
- Validating dose reduced imaging techniques 13
- Software solutions for facilitating observer performance studies 14
- Internet technologies 15
- Continuous variables 16

Aims 17

Materials and Methods 19
- Design of the web-based platform 19
  - Database tier 21
  - Middle tier 21
  - Client tier 22
- The Mulrecon DICOM viewer 23
- Updated Mulrecon DICOM viewer 26
- Web security and web server hosting 26
- Usability and application performance 26
- Clinical application studies 29
  - Patient populations and imaging studies 32
  - CT low dose simulations 33
  - Case assessments using the WOAP 34
  - Statistical analysis 37

Results 43

Discussion 49
- Emergent proposals for web-based facilitation of observer performance studies 51
- Implications for use in clinical research studies 53
- Limitations 55
- Future perspectives 56

Conclusion **57**

References **59**

Appendix A **70**

Appendix B **71**

Appendix C **73**

Thesis papers I-IV **75**

List of Figures

1.1 A volumetric dataset may be visualized… 10

1.2 Example of processing of a volumetric CT data set… 11

1.3 CT multiplanar reconstruction depicting the abdominal aorta… 12

2.1 Schematic overview of how clinical application studies 1-3 and the method study… 18

3.1 Screenshot showing the web-based instruction page for study 1… 22

3.2 Screenshot from the measurement module of the web-based platform… 23

3.3 Screen capture of the Mulrecon interface with a stack of thorax… 24

3.4 The System Usability Scale… 28

3.5 Schematic transverse image of the abdominal aorta… 32

3.6 The low-dose simulation technique used in study 2… 33

3.7 The web-based DICOM viewer used in study 2… 35

3.8 Pictorial presentation of two-point caliper placement… 36

3.9 Agreement plots for each of the three methods (OTO, LTL, and ITI)... 38

3.10 The interactive web-based statistical module with an agreement plot… 41

Introduction

This Ph.D. thesis concerns the development and application of a web-based platform for the facilitation of observer performance in medical imaging. But before we dive into the details of observer performance in relation to imaging, the central terms *imaging biomarker*, and *imaging efficacy* will be outlined to the reader.

The past couple of decades have seen medical imaging capabilities dramatically expand. Modern techniques, including ultrasound, computed tomography (CT), magnetic resonance imaging (MRI), and positron emission tomography (PET), now provide an abundance of data and an unprecedented level of spatial detail and functional information[1]. The term biomarker has increasingly been associated with diagnostic imaging, and a biomarker can be defined as "*any medical sign or characteristic that objectively measures a normal or pathological process or a response to treatmen*t"[2]. Of late, a significant emphasis has been focused on quantitative imaging biomarkers[3,4,5]. However, all biological characteristics detectable in an image are essentially biomarkers. Along these lines, imaging biomarkers can broadly be divided into qualitative and quantitative, and it follows that qualitative biomarkers are those that cannot be expressed using quantitative values - e.g., descriptive; "*a nodule is present in the lung*," and pathological grading systems such as the Prostate Imaging Reporting & Data System. Conversely, quantifiable biomarkers are those whose magnitude is expressed in numbers (e.g., diameter, volume, density, intensity diffusion, and variables from positron emission tomography such as standardized uptake value, etc.)[5]. Faced with constant technological innovation, it is imperative to assess the value of new potential imaging biomarkers. There is an increased societal demand for evidence that imaging (biomarkers) affects patient outcomes, and its cost burden on the healthcare system is questioned[6]. In this context, the term efficacy is helpful.

Medical imaging efficacy

In a seminal paper from 1991 - The efficacy of Diagnostic Imaging - Fryback and Thornbury defined efficacy as: "*the probability of benefit to individuals in a defined population from a medical technology applied for a given medical problem under ideal conditions of use.*"[7]. The authors outlined a hierarchical model for classifying the scientific evidence of imaging efficacy (Table 1.1). This model consists of six different levels ranging from the most straightforward foundation of technical aspects of image capture, such as image resolution (Level I), up to the highest level of efficacy, which they

referred to as “societal efficacy” (Level VI).
Fryback and Thornbury emphasized that for an imaging procedure/biomarker to be efficacious at a higher level in this hierarchy, it must be efficacious at lower levels, but the reverse is not true.
Concerning the diagnostic accuracy of the imaging system (Level II) and the statistical evaluation of the performance of diagnostic imaging, Fryback and Thornbury remarked, “*Important to [note about] all these measures is that they attempt to measure performance of the imaging for the purpose of making diagnoses and that they all require interpretation of the imaging by an observer….diagnostic accuracy efficacy is not simply a function of the image. It is a joint function of the images and of an observer, such as a radiologist.*”[7]. This remark brings us back to the core of this Ph.D. thesis, namely observer performance in imaging research and development.
Assessment of observer performance is an essential element in the evaluation of diagnostic accuracy as well as observer agreement in medical imaging. However, in terms of definitions, it is important to emphasize that agreement studies focus on the variability of evaluations performed by different observers on the same subjects without requirement of a reference standard[8]. In clinical practice, imaging is increasingly used when a clear reference standard is not available, and in such circumstances, agreement studies are used to assess the objectivity of imaging results. This focus diverges from studies of diagnostic accuracy in which obtained measurements/results are compared with a reference standard (known truth)[9]. Nevertheless, it is not possible to have a highly accurate imaging test that is subject to significant observer variability. Hence, a preclude to substantiating that an imaging test has high accuracy, is documenting sufficient observer agreement.

Level of Efficacy	**Examples of Endpoints for Each Level of Efficacy**
Level I: technical efficacy	Imaging resolution
Level II: diagnostic accuracy efficacy	Test sensitivity/specificity. Reproducibility.
Level III: diagnostic thinking efficacy	Pre- and posttest changes in subjectively determined outcome
Level IV: therapeutic efficacy	Effects of diagnostic on the choice of therapy
Level V: patient outcome efficacy	Value of test information, including measures of morbidity, mortality, and quality-adjusted life years.
Level VI: societal efficacy	Cost-benefit and cost-effectiveness from the societal perspective

Table 1.1: Six levels of efficacy and challenges for comparative effectiveness research[7].

Observer performance studies in imaging research

Observer performance studies in imaging are typically conducted in a setting where observers read imaging cases at a particular time and place using a PACS (Picture Archiving and Communication System) workstation or DICOM (Digital Imaging and Communications in Medicine) viewer installed on a standalone personal computer. The DICOM file format is a universally standard for medical imaging storage and transmission adopted by virtually all manufacturers. Case assessment results are recorded in an electronic database or with pen and paper. This practice presents logistical challenges, and what follows is an account of the need for measures to facilitate such studies and lower the threshold for observer participation. In an editorial published in 1991, Beam et al. discussed the problem of establishing level II efficacy and highlighted the need for diversity of observers in imaging research: "*When can radiology research be extrapolated to the whole profession? The answer is 'never,' as long as studies in diagnostic radiology continue to utilize only a small number of highly selected readers who represent expertise only at the upper level of their subspeciality.*"[10]. Given the preceding arguments, Beam et. listed five essential questions which, unfortunately, are likely to be unanswered for most imaging techniques:

1. How much does an imaging technique improve the diagnostic ability of the average radiologist?
2. How much of an improvement over the use of a reference technique will this new technique typically make?
3. How much variability in diagnostic abilities is there to be found in the general population of radiologists and in subspecialties?
4. Is gain in diagnostic performance dependent on characteristics of the radiologists (e.g., years of experience, specialty, training) and if so, how?
5. How much disagreement in diagnosis is to be naturally expected between radiologists using the same imaging technique or for the typical radiologist when reinterpreting the same images?

In addition, an interesting and illustrative analogy was put forth by Beam et al., "*imaging techniques are akin to 'treatments' that we apply to subjects (radiologists) and the response we measure in these subjects are diagnostic success rates.*"[10]. In a 1996 perspective paper, Obuchowski and Zepp have expanded upon this analogy, "*Consider how little we would learn about a 'treatment' if the study had only one subject (single-reader study) or if the study had multiple subjects but their individual responses were not recorded but rather were expressed as a single pooled response (consensus*

reading study)."[11]. Obuchowski and Zepp argue that imaging research has seen an "*inappropriate focus on the accuracy of the imaging system rather than on the accuracy of the readers interpreting the images.*" To provide a scheme for assessing level II efficacy and striving to answer the above five questions, Obuchowsky and colleagues defined different types of studies needed for such an endeavor which are also arranged in a hierarchy (Table 1.2)[12,13].

Phase I (exploratory/preliminary study)	Phase II (challenge phase)	Phase III (advanced phase/mature tests)
Purpose: To determine whether the test can distinguish between those with clear disease and healthy subjects. Should be restricted to preliminary investigations. Patient sample size: 10-50 Observers: 2-3	Purpose: To compare the accuracy of the tests and examine the relationship between accuracy and the pathologic, clinical, and comorbid features. Patient sample size: 50-200 Observers: 5-10	Purpose: Estimate the performance of the imaging system for a well-defined population of patients and a well-defined population of observers who will use the medical device Observers: > 10

Table 1.2: Types of studies for assessment of diagnostic performance arranged hierarchically[12].

The recommendation Obuchowsky and colleagues provide regarding proper patient sample size and the number of observers that ought to be included in such studies mainly pertains to diagnostic accuracy studies involving categorical and ordinal variables evaluated with statistical measures such as receiver operating characteristics (ROC). Nevertheless, these recommendations are equally relevant concerning studies that evaluate observer agreement in, for example, size- and volume-based imaging measurements involving continuous variables. It has been underlined that it is an erroneous assumption that such measurements of continuous variables are obtained through an objective process without uncertainty[14]. For example, it is known that CT-based tumor sizing based on manual measurements can be subject to substantial intra- and interobserver variability. This is especially the case in the presence of irregular tumor margins, which might only be exposed when a variety of observers perform measurements[15,16,17].

In terms of research practice, two papers have investigated the actual number of observers employed in observer agreement studies. Shiraishi and colleagues analyzed ROC studies in the journal Radiology between 1997 and 2006[18]. They found that nearly 50% of studies included three or fewer observers and concluded that this practice seriously challenges the generalizability of their conclusions to the relevant population of radiologists. Farzin et al. recently sought to estimate the frequency and quality of observer agreement studies published in four imaging journals, including Radiology, between 2011 and 2012. Of

2229 studies, 280 (13%) evaluated agreement, and in 81% of these studies, two or fewer observers were included[19]. Farzin et al. concluded that agreement studies are infrequently reported and that the number of observers included in such studies was small. In addition, they concluded that most investigations should be regarded as explorative and that agreement studies are research opportunities that should be promoted.
It is clear that the numbers provided by Farzin et al. and Shiraishi et al. are in contrast to the numbers outlined in Table 1.2 as recommended by Obuchowsky et al.[11].

Based on past editorial experience, Bankier et al. have speculated as to the reasons why researchers fail to report variability between observers and techniques with a sufficiently large number of observers included[20]. The first reason is the relative ease with which consensus readings are performed. The second reason is the above-mentioned inappropriate focus on the imaging system rather than on the observers. Bankier et al. highlight that variations between observers are determined by several factors such as technical skills, perceptive skills, training, and experience. It needs to be recognized that many authors appear to perceive such observer differences as detrimental because they can make imaging techniques look less advantageous once variability is reported[21]. In addition, a third reason might be a lack of familiarity with the statistical tools designed for this purpose. Concerning the statistical evaluation of agreement in continuous variables, Jones et al. have observed that very few studies include more than two observers and speculated that one reason is the fact that the widely used Bland-Altman methodology only accommodates two observers[22]. Finally, a fourth reason proposed by Bankier et al. is the cost and above-mentioned logistical challenges in obtaining a large sample of observers. Another significant aspect that must be taken into account because it adds to the difficulties in conducting observer performance studies is the widespread use of volumetric imaging datasets.

The volumetric imaging revolution

Volumetric medical images, such as CT and MRI scans, are composed of a series of stacked two-dimensional images (Fig. 1.1-1.2). Volumetric imaging was introduced into medical practice in the 1970s, and from around 2005, the radiological practice has seen a dramatically increased usage of volumetric images[23]. The emergence of imaging processing capabilities in PACS environments allows real-time manipulation of these volumetric datasets. Consequently, more complex and time-consuming human-computer interactions such as scrolling, alteration of window settings, and on-the-fly multiplanar reconstructions (MPR) and volume rendering with 3D models have become an integral

part of contemporary radiological practice (Fig. 1.3)[24]. In fact, it has been demonstrated that the cognitive processes of the radiologist involved in volumetric image interpretation differ substantially compared to a 2D paradigm[25]. Regarding observer performance studies, this means that such studies need to be conducted with a radiology-workstation-like interface where the commonly found image manipulation tools are available to mimic daily radiological clinical practice. Taken together, this complexity of diagnostic imaging and the resulting increased time-consumption of image interpretation adds to the difficulty in recruiting a sufficient number and variety of observers willing to commit the time required for participation.

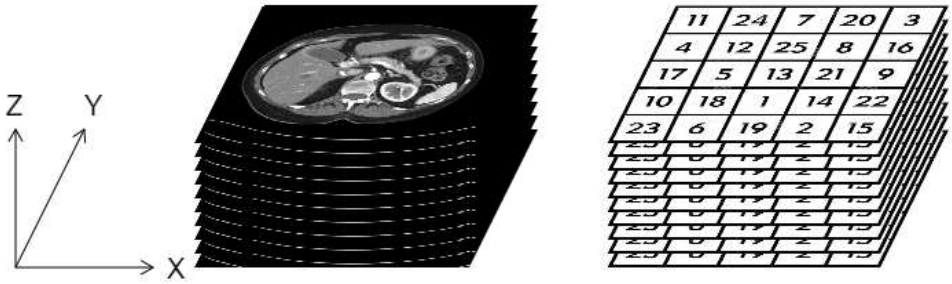


Fig. 1.1: A volumetric dataset may be visualized as a stack of several hundred to more than one thousand 2D axial image sections obtained with, for example, MRI or CT at regular intervals along the z-axis. Typically each image has dimensions of 512x512 pixels. The 2D images are transformed into a 3D dataset composed of voxels. A voxel describes the dimensions and position of the smallest cube in a dataset. The voxels are arranged in a Cartesian volume, each associated with an x-y-z position and an intensity value. For illustrative purposes, the dimensions of the images in this figure are 5x5 pixels (Borgbjerg 2021).

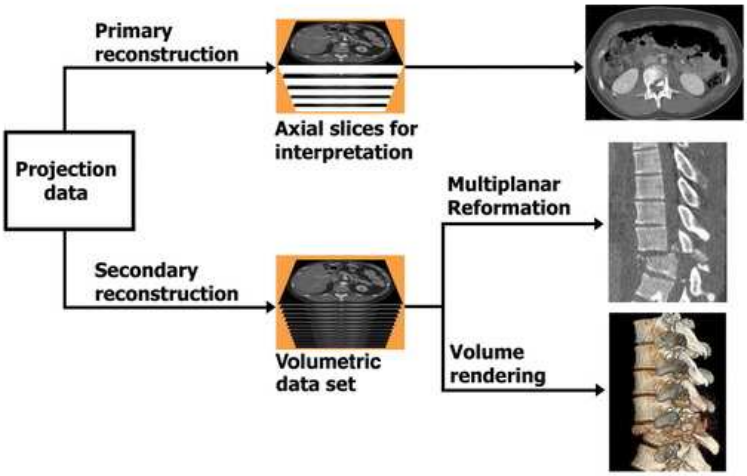


Fig. 1.2: Example of processing of a volumetric CT data set. CT data are typically used to reconstruct axial images of interpretive thickness for conventional review, which is performed using a picture archiving and communication system. If necessary, a thin-section data set can be generated in addition to or in place of the traditional interpretive axial images. This may be called the volumetric data set because it is intended not for primary axial interpretation but rather for generating high-quality multiplanar reformatted or volume-rendered images. This data set typically consists of axial images with a section thickness approaching 1 mm or even less, preferably with an overlapping interval. Reprinted and adapted from Radiographics 2005[26], copyright, by permission of Radiological Society of North America

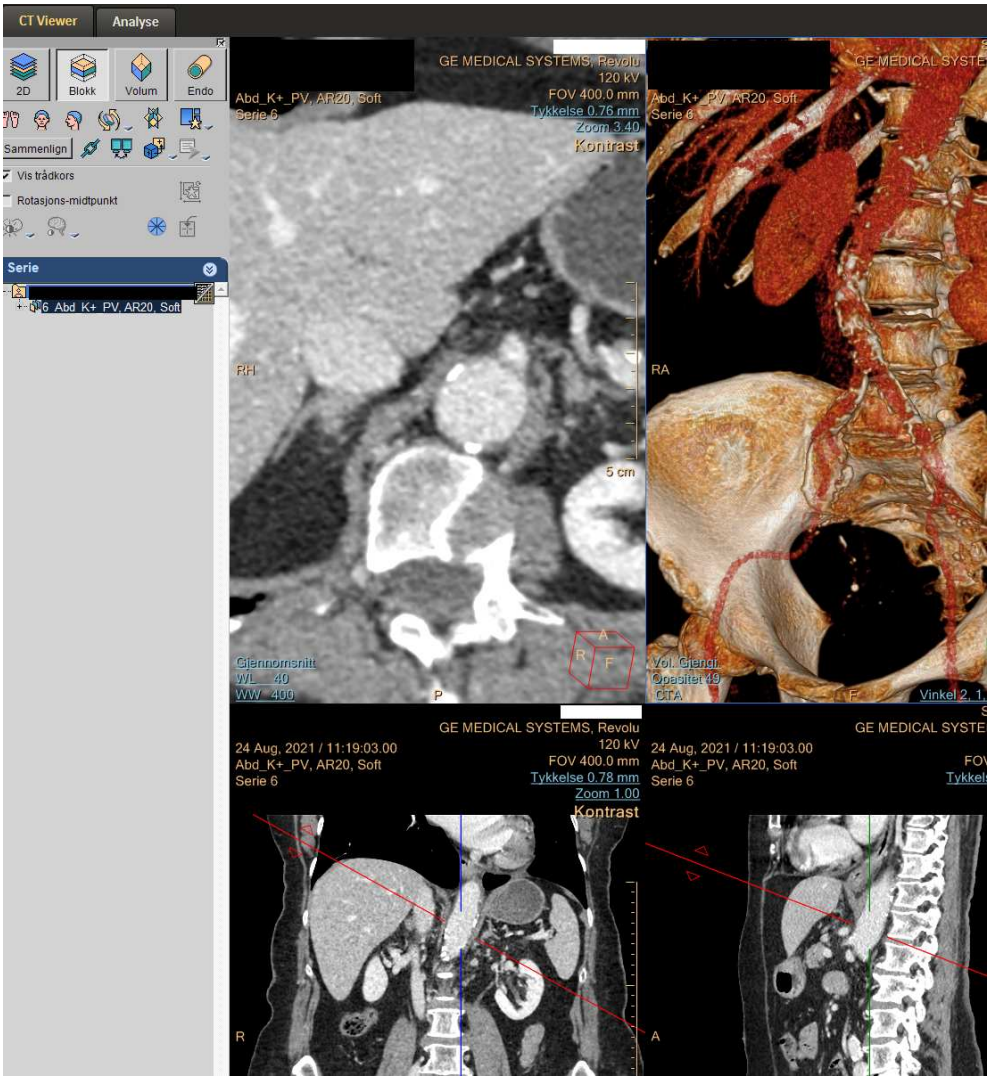


Fig. 1.3: CT multiplanar reconstruction depicting the abdominal aorta performed with the Philips IntelliSpace Portal (Philips Medical Systems). In addition, volume rendering is shown in the upper right corner (Borgbjerg 2022).

Validating dose reduced imaging techniques

One area in particular in which there is a pressing need to assess observer performance is in relation to evaluating strategies for reducing radiation dose in medical imaging[27]. According to the European Directive Euratom, all member states of the European Union need to ensure justification and optimization of radiological procedures and store information on patient exposure for analysis and quality assurance[28]. The dose principle of As Low As Reasonably Achievable (ALARA), consistent with the diagnostic task, is advocated by radiological societies[29]. In adherence to this principle, it is imperative to evaluate whether a different modality such as MRI or a lower dose CT scan can replace a conventional higher dose CT scan. Technological advancements in CT continue to provide many new avenues for radiation dose optimization[30], and it is well-recognized that a large variability exists between the doses needed for different diagnostic tasks[31]. Hence, radiation dose should be tailored not simply to the body part being imaged but rather for the diagnostic task in question. For example, for CT-based detection of low-contrast liver lesions, evaluation is compromised with modest radiation dose reduction[32]. In contrast, a 100-fold radiation dose reduction is feasible for torsion measurements of the lower limb[33]. To determine if a scan based on a low-dose CT protocol is a satisfactory substitute for a higher dose scan for a given diagnostic task, one has traditionally needed to conduct repeated scans of patients resulting in a significant increase in radiation exposure. Fortunately, techniques that allow the generation of simulated low-dose CT images from an original higher dose scan have become available[30]. Data from a feasibility study by Fletcher et al. examining a range of simulated dose levels for common CT examinations suggest that an opportunity exists for substantial dose reduction using existing CT technology[34]. In addition, Fletcher et al. demonstrated that radiologists' subjective confidence in diagnostic image quality generally declines before measures of observer performance. This finding underscores the need for assessing observer performance in a setup evaluating the relevant diagnostic task and not simply a surrogate marker such as image quality assessment. These new options for low-dose simulation do not solve the logistical challenges as part of conducting clinically relevant observer performance studies. Despite the continual emergence of new potential imaging biomarkers and dose-saving strategies, it is also important to reiterate that imaging measurements such as tumor size used in cancer staging and assessment of treatment response for decades have, in some cases, never been adequately evaluated in terms of observer variability[35].

Software solutions for facilitating observer performance studies

Existing PACS systems are closed source systems and very difficult to customize. In general, they do not accommodate workflow automation in performance studies, and study data cannot be saved automatically. At the beginning of the research work for this thesis (2016), only a few software solutions specifically tailored for the facilitation of observer performance studies were described in the literature. One notable example is ViewDex, an abbreviation meaning "viewer for digital evaluation of x-ray images," first reported in 2005[36]. It is a free-to-download desktop-based solution written in Java, DICOM compatible, and incorporates automatic data collection. The software was initially developed for visual grading of static x-ray images but has undergone continuous development and can now display stacks of cross-sectional images, but does not support image processing techniques such as MPR described above. More significantly, it does not allow observers to complete study readings from different sites. An alternative to VievDex is MedXViewer which was presented in 2016 and was primarily developed for use in observer performance studies evaluating digital mammography and tomosynthesis[37]. The software is also written in Java and desktop-based. It provides similar functionality to ViewDex and does not feature more advanced capabilities such as MPR of volumetric datasets. However, one notable difference compared to ViewDex is an option to integrate MedXViewer with a web-based database. The web-based database allows for central storage of imaging studies to be evaluated and the results of case assessments. Thus, observers can participate from different locations, albeit it still requires the local installation of software and storage of cases to be evaluated. This represents a barrier to the recruitment of participating observers. Additionally, the software is not available for direct download.

An alternative to a desktop-based solution is a platform based entirely on the Internet. Listed in Table 1.3 are a number of the potential advantages compared to a "pen and paper" approach that can be harnessed using the Internet for observer performance studies.

Avoiding pen and paper or data entry in an external database allows for observers to focus mentally and physically on case readings
The risk of making mistakes in registering the results of observers' evaluations is mitigated
More extensive data, including key images, can be saved from case readings
A web-based platform allows for the silent deployment of data and code updates
Accessibility and availability; any user using a web browser can participate without installing additional software.
Diagnostic image displays from radiological workstations can be used. Installing and using external programs in a hospital environment is often not possible due to cybersecurity issues or will at least require the involvement of a network technician.
Increased flexibility where case readings can be performed by observers scattered in time and place
A potentially unlimited number of cases can be evaluated, and a large number of observers can be recruited, which would not be possible in a lab-based setting

Table 1.3: Potential advantages of conducting observer agreement studies on an internet platform.

Internet technologies

Being a very vision- and technologically-driven field, one would assume that diagnostic imaging would quickly utilize the Internet and a web-based model for experimentation involving human observers. Nevertheless, for example, compared to psychological research, imaging research has been slow to harness web-based experimentation. Web-based behavioral research beyond simple questionnaires in, for example, spatial cognition and visual motion can be dated to the late 90s[38]. One of the reasons for the lag concerning imaging research is the capabilities of web browsers of the 2000s. Back then, the execution speed and data transfer of dedicated PACS networks and computer workstations were adequate for displaying volumetric datasets; however, web browsers had speed and memory limitations and were inconsistent in displaying interactive graphics[39]. Java and Flash are installable plugins enabling advanced graphics manipulation to be initiated in a web browser widely used in the 2000s. However, plugin-based web applications leave the host computer vulnerable to cyber-attacks and have lost browser support[40]. Fortunately, the advent of the Hyper Text Markup Language 5 (HTML5) canvas element in 2014 paved the way for platform-independent pixel-level manipulation and display of data in the web browser without the need for plugin installation. Furthermore, the canvas element was paired with the JavaScript API WebGL (Web Graphics Library). WebGL allows hardware-accelerated graphics in the web

browser by providing access to the Graphics Processing Unit (GPU) such that rendering speed approaches desktop applications. In a proof of concept study from 2011, using WebGL, Congote et al. exhibited volume rendering in the web browser[41], which demonstrated the feasibility of implementing DICOM viewers accommodating volumetric datasets in a web application.

Continuous variables

As already highlighted, even the seemingly "objective" imaging assessment of uni-dimensional size can be subject to substantial observer variability. One area in which there is a strong correlation between a uni-dimensional imaging biomarker and patient outcome is abdominal aortic diameter. The diameter of an abdominal aortic aneurysm (AAA) is a strong predictor of rupture risk and plays a prominent role in AAA management. Population-based screening using ultrasound has been established in several countries, and in addition, AAAs are detected regularly as incidental findings when cross-sectional imaging is performed for other indications. Reproducible measurements of abdominal aortic diameter are of paramount importance because measurement imprecision can negatively affect care pathways in AAA management; for example, inappropriate enrolment into surveillance programs at the 30 mm threshold, delayed surgical referral at the 55 mm threshold, or lack of recognition of expanding AAA after endovascular aneurysm repair[42]. The accuracy and reproducibility of ultrasound assessment of maximum aortic diameter involve several factors contributing to measurement variance. These include operator skill and training, ultrasound machine settings and frequency, the habitus of the patient, degree of intimal plaque calcification, presence of mural thrombus, aortic curvature, the plane of image acquisition, the axis of measurement, diameter selection, aortic level, cardiac cycle, and caliper placement. A number of these factors also apply to other cross-sectional imaging modalities such as CT and MRI. It has yet to be determined what role these can play in potential opportunistic screening and rationalized systematic screening[42,43]. Another area in which the potential role of uni-dimensional size measurements is currently being debated is the diagnosis, grading, and follow-up of chronic pancreatitis patients[44]. Aspects of observer agreement in diameter assessments in patients with abdominal aortic aneurysm and chronic pancreatitis will be explored in this thesis.

Aims

The overall aim of this Ph.D. thesis was to develop and apply a web-based platform for facilitating observer agreement studies in clinical imaging research and development. A particular focus was on evaluating the measurement variability of uni-dimensional continuous variables. As part of the Ph.D. thesis, three clinical application studies were completed (Studies 1-3), as well as one method study. Consequently, the thesis is based on one method paper (II) and three application papers (I, III-IV), in which, during the completion of the latter three, the platform was iteratively improved (see Fig. 2.1).

To fulfill the overall aim, the thesis contains six specific aims:

I. To develop and improve a web-based database application for observer authentication, study management, and storage of evaluation results (I, III, IV)
II. To determine reproducibility using the web-based application in the ultrasonographic assessment of maximum abdominal aortic diameter with three principal methods of caliper placement (I)
III. To develop an improved and easily accessible web-based DICOM viewer for visualization and manipulation of volumetric datasets (II)
IV. To assess the interchangeability of ultra-low-dose non-contrast CT and standard-dose CT angiography using the improved web-based application in terms of accuracy and reproducibility in determining maximum abdominal aortic diameter. (III)
V. To quantify the level of observer variability using the improved web-based application in CT-based measurements of ductal- and gland diameters in chronic pancreatitis (IV)
VI. To assess the overall usability and performance of the developed platform (I-IV)

In fulfillment of thesis aims I and VI, we report data not explicitly reported in I-IV but nonetheless acquired as part of studies 1-3 conducted for I and III-IV.

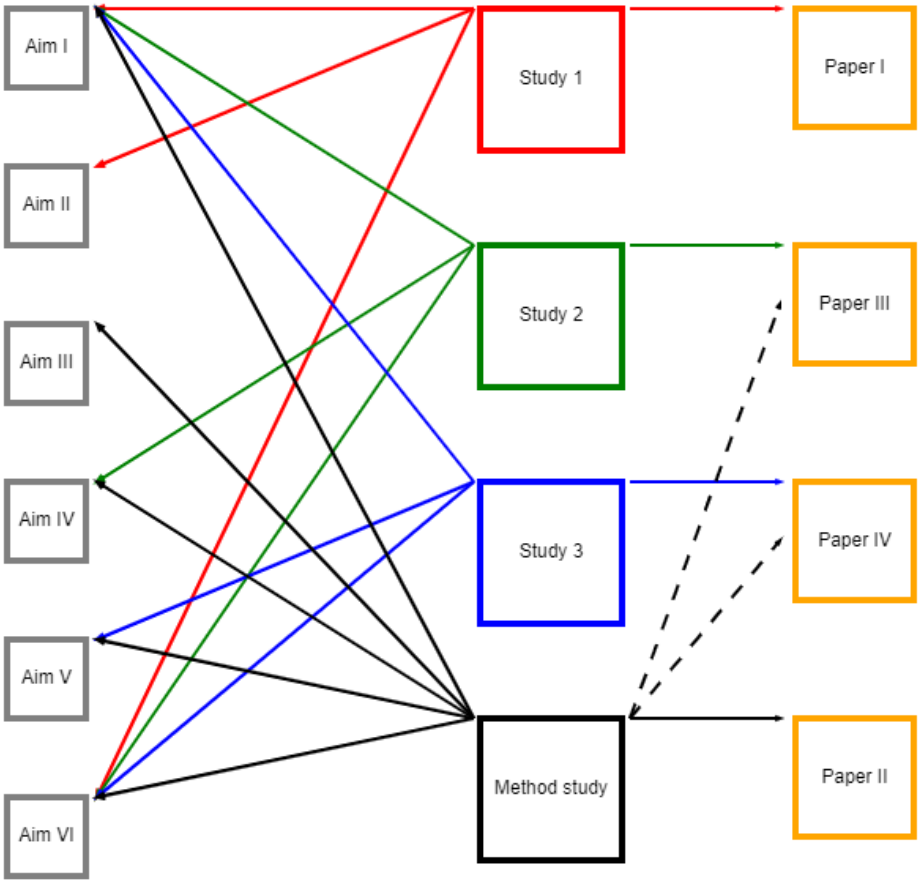


Fig. 2.1: Schematic overview of how clinical application studies 1-3 and the method study relates to the resulting papers I-IV and thesis aims I-VI. Colored lines between studies and aims/papers signifies how the study in question contributes to the aims/papers in question. Dashed lines signifies that the volumetric DICOM viewer developed as part of the method study did not directly lead to papers III-IV, but nevertheless had a central role in the completion of clinical application studies 2 and 3 reported in papers III-IV (Borgbjerg 2022).

Materials and Methods

In the following sections, I detail the platform's development process for the facilitation of observer performance studies (henceforth, the developed web-based platform is referred to as WOAP).
Subsequently, the three application studies are described.

Design of the web-based platform

Primary goals in the design of the WOAP were modularity, expandability, and accessibility.
A variety of web frameworks exist today for the development of web-based platforms. A PHP-MySQL combination for the data management system was chosen due to the open-source nature of PHP (PHP: Hypertext Preprocessor) because it is widely supported on web servers and has a long history of reliability[45]. The platform is based on the well-known three-tier web architecture, as shown in Table 3.1[46]. At the base of WOAP is the database tier, consisting of the database management system MySQL, which uses the SQL (Structured Query Language) language for adding, accessing, and managing the contents in the database. The middle tier is built on top of the database tier, which contains the web server that stores downloadable files. This tier also communicates data between the other tiers, achieved through PHP scripting. On top is the client tier, which is the web browser that receives the HTML5 (Hypertext Markup Language revision 5) and Javascript code needed to implement and present DICOM viewer functionality. The middle tier implements a web interface for the semiautomatic setup of the parameters for a new study. The study administrator can further configure these with the phpMyAdmin, a web-based administration tool for MySQL. What follows is a further specification of the three tiers of the WOAP.

Client tier	Front end	The client program (Web browser)	HTML, JavaScript, WebGL
Middle tier	Back end	Webserver	PHP
Database tier		Database management system	MySQL

Table 3.1: The three-tier architecture of the web-based platform.

Setting up a study first entails using a web-based administrator module to copy the basic PHP files of the WOAP to a new folder on the webserver and create an associated database. Table 3.2 lists the principal steps one goes through as a participant in an agreement study with the WOAP.

1.	Observer invited
2.	Observer accepts to participate in the study
3.	Observer receives login information to the WOAP
4.	Observer signs off on case assessment instruction
5.	Observer practices using the imaging viewer
6.	Observer completes 1st session of case assessments
7.	Observer completes 2nd session of case assessments after a designated waiting period
8.	…
9.	Observer completes Xth. session of case assessments after a designated waiting period
10.	Observer has finished case assessments
11.	Observer receives feedback from study administrator (if applicable)

Table 3.2: Principal steps one goes through as a participant in an observer performance study when using the WOAP..

Database tier

The original structure of a study database template is largely unchanged from the initial version of WOAP until this writing. This database is composed of four tables, as seen in Table 3.3.

Cases table	Manages the imaging cases that need to be evaluated for a given study.
Observers table	Lists and manages participating observers and specifies observer authentication credentials. The table allows the study administrator to follow and change the study status of observers (e.g., after a sufficient waiting period enable access for an observer to make a new round of repeated measurements) and monitor case assessment progress (e.g., how many cases are left in a given round)
Measurements table	Lists the measurements to be completed by each observer. This table registers the start and end times of measurements. A JSON format stores registered parameters for each case assessment (e.g., size/distance/angle measurements, visual grading scale, free comments, x-y-z coordinates, screenshots, etc.)
Study parameters table	This table contains headers to specify overall study parameters (e.g., study instructions, number of cases, number of repeated case assessments, and which measurements/evaluations must be made for each case)

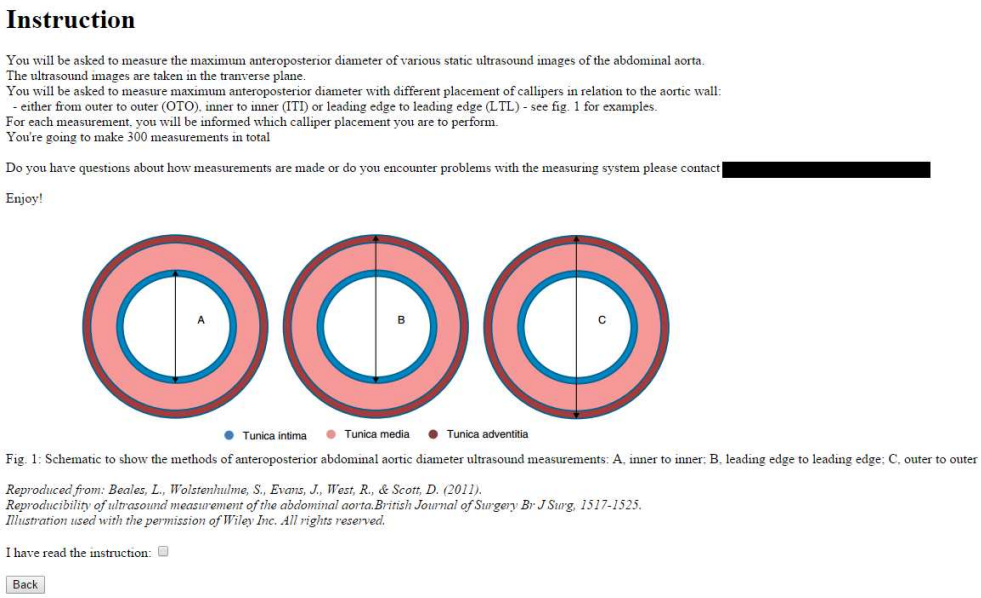
Table 3.3: The basic table structure of the WOAP.

As mentioned above, the phpMyAdmin can alter the contents of these tables, including adding new table headers, without altering the basic data flow. In the study parameters table, a header defines a JSON (JavaScript Object Notation) file that specifies a list of the measurements/evaluations to be completed for each case assessment. JSON is an open standard file and data interchange format that uses human-readable text to store and transmit data objects. When presenting a case to a given observer for assessment, the DICOM viewer retrieves the JSON file, which interprets the file and prompts the observer to complete the specified case assignments. Subsequently, the results of these assignments are used to generate a JSON file and, upon case completion, saved in the relevant entry of the measurements table.

Middle tier

PHP generates the HMTL5 and Javascript code sent to the client-tier to present a web-based observer login system as well as an instruction step that prompts each user to review case assessment instructions before performing case assessments (Fig. 3.1). The web server stores the DICOM files associated with each case specified in the database

cases table. For a given study, following manual patient de-identification, selected DICOM studies can be uploaded to the web server using a File Transfer Protocol (FTP)-client or by a web-based upload system.



Instruction

You will be asked to measure the maximum anteroposterior diameter of various static ultrasound images of the abdominal aorta.
The ultrasound images are taken in the tranverse plane.
You will be asked to measure maximum anteroposterior diameter with different placement of callipers in relation to the aortic wall:
- either from outer to outer (OTO), inner to inner (ITI) or leading edge to leading edge (LTL) - see fig. 1 for examples.
For each measurement, you will be informed which calliper placement you are to perform.
You're going to make 300 measurements in total

Do you have questions about how measurements are made or do you encounter problems with the measuring system please contact [redacted]

Enjoy!

Fig. 1: Schematic to show the methods of anteroposterior abdominal aortic diameter ultrasound measurements: A, inner to inner; B, leading edge to leading edge; C, outer to outer

Reproduced from: Beales, L., Wolstenhulme, S., Evans, J., West, R., & Scott, D. (2011).
Reproducibility of ultrasound measurement of the abdominal aorta.British Journal of Surgery Br J Surg, 1517-1525.
Illustration used with the permission of Wiley Inc. All rights reserved.

I have read the instruction: ☐

Back

Fig. 3.1: Screenshot showing the web-based instruction page for study 1 (Borgbjerg 2022).

Client tier

The imaging viewer was written as a single-page application[47]. Hence, the browser retrieves all necessary HTML and JavaScript code with a single HTML page load in this approach. For the study reported in I, an initial version of the viewer was built, which could only accommodate single images in JPEG format exported from DICOM files. This viewer only implemented a caliper measurement functionality where no other 2D image manipulation tools such as zoom, contrast/brightness adjustments, pan, etc., were made available (Fig. 3.2). Later, the Mulrecon DICOM viewer reported in II was developed and integrated with the improved WOAP used in III and IV. The Mulrecon viewer is described in greater detail below. With respect to loading cases for assessment, the same principle was used for both the initial viewer and the Mulrecon viewer: based on a reference provided by the middle tier, the appropriate resources (i.e., DICOM files and JSON file

case assessment specification) were dynamically retrieved from the middle tier and added to the viewer. Consequently, the viewer represents an independent module where the code does not need to be customized when setting up a new study.

Fig. 3.2: Screenshot from the measurement module of the web-based platform employed in I showing caliper placement in a transverse ultrasound image of the infrarenal abdominal aorta using the outer to outer method.
Reprinted from European Journal of Vascular and Endovascular Surgery 2018[48], copyright, by permission of Elsevier.

The Mulrecon DICOM viewer

A single-page web application was developed to mimic a DICOM viewer for visualization of volumetric datasets (Fig. 3.3). The initial version, as presented in II, was written entirely in JavaScript. Based on Javascript and HTML5, it can provide a unified user experience across web browsers[47]. Several JavaScript frameworks were used to implement functionality and a graphical user interface (Table 3.4).

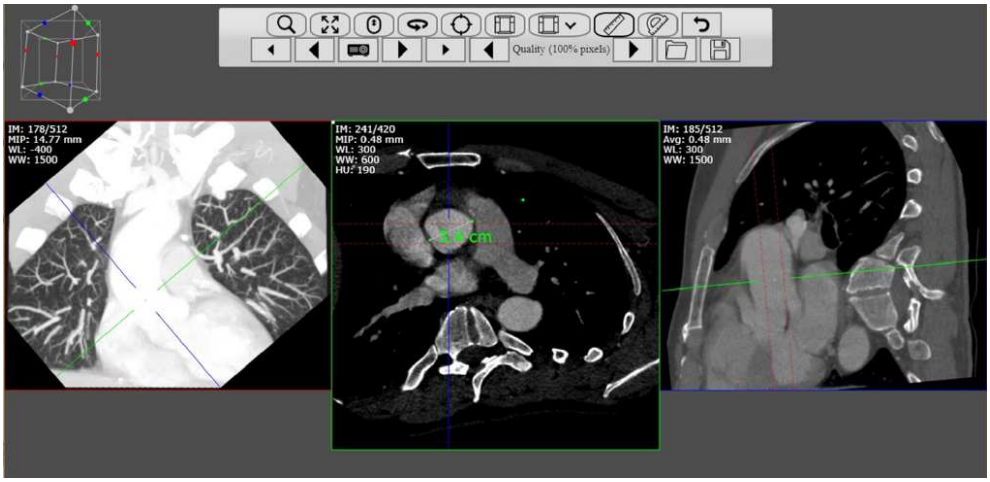


Fig. 3.3: Screen capture of the Mulrecon interface with a stack of thorax CT DICOM images as input (II). Double oblique multiplanar reconstructions have been rendered with measurement of ascending aortic diameter. Thick slab maximum intensity projection is rendered in a lung window as well.
Reprinted from Current Problems in Diagnostic Radiology 2019[49], copyright, by permission of Elsevier.

Most notably, the dicomParser.js framework was used to parse DICOM files. Each DICOM file has a header containing patient demographic information, acquisition parameters, study modality, image dimensions, and pixel data[50]. DICOM tags were used to identify and display the type of study in question correctly. Custom JavaScripts were implemented to arrange loaded slices based on their orientation (i.e., axial, sagittal, etc.) and position in space as determined from DICOM attributes. Subsequently, a 3D scalar field of voxel values is generated. The viewer implements multiplanar reconstructions (MPR). One-voxel-thick sections of the volumetric dataset can be generated to display the standard imaging planes (i.e., axial, sagittal, and coronal) as well as oblique reconstructions. MPR sections can also be generated with different slice thicknesses by projecting and sampling the dataset along lines (ray casting) perpendicular to the imaging plane within the desired display area. The three standard image planes are initially displayed with cross-reference lines when the viewer has loaded a dataset. Subsequently, these imaging planes can be manipulated as in a typical PACS (Table 3.5), including the generation of double oblique reconstructions as seen in, for example, visualization of the aorta perpendicular to the flow direction. A trilinear interpolation algorithm written in JavaScript was used to generate smooth reconstructions. The trilinear interpolation algorithm is the most popular algorithm for volumetric reconstruction but is

computationally expensive[51]. JavaScript Web workers were deployed for parallel processing using multiple central processing unit (CPU) cores to increase rendering speed. However, despite the utilization of multiple cores, the implemented software-accelerated rendering speed was still insufficient; smooth reconstructions could only be generated with latency, even for one-voxel-thick sections, when a user ceases to manipulate the dataset.

Framework	Functionality
dicmParser.js	Used to read DICOM files
statjs	Implementations of statistical functions
mathjs	Implementations of mathematical functions
jQuery	A library designed to simplify HTML element manipulation
jQuery UI	A collection of standard graphical user interface elements such as dialog boxes

Table 3.4. Selected frameworks used for the Mulrecon DICOM viewer.

Function
Pan Zoom Scroll Synchronize Caliper, polygon, and angle tool Slice thickness alteration Projection technique (MPR, MIP, MinIP) Rotation Save images Window-level alteration

Table 3.5: List of Mulrecon functionalities as presented in II

Updated Mulrecon DICOM viewer

As part of setting up the study reported in III, display requirements for volumetric imaging data in medical viewing applications as specified by the Royal College of Radiologists were reviewed and implemented as an extension of the Mulrecon viewer first presented in II[52]. The viewer was updated with hardware-accelerated rendering based on WebGL[53]. As previously outlined, WebGL enables access to the client device's graphics processing unit (GPU). Briefly, the GPU is a processor made up of many smaller and more specialized cores that can deliver massive performance when a processing task can be divided up and processed across many cores[54]. Hence, WebGL was used to increase application speed by executing the time-consuming ray casting computations for each pixel in the rendered image planes in parallel. WebGL version 2.0, introduced in 2017, was used, which in contrast to WebGL 1.0, enables trilinear interpolation algorithm directly supported in the hardware. The implemented hardware-accelerated rendering allows the updated viewer to perform real-time smooth reconstructions compliant with modern radiological practice. Additionally, hotspots were created to select zoom, pan, synchronize, and rotation functions to facilitate easier manipulation of image stacks. A working version of the Mulrecon DICOM viewer with sample volumetric datasets available can be found online (https://www.castlemountain.dk/atlas/index.php?page=mulrecon&mulreconPage=color).

Web security and web server hosting

The open-source nature of the combination of PHP-MySQL allows the WOAP to run on a plethora of web servers. For the studies reported in I, III-IV, a commercial web host was chosen with an annual price of approximately 150 euros. The server has a Secure Socket Layer (SSL) certificate. A Secure Socket Layer is a standard security technology for encrypting the connection between the client user and the web server[55]. Hence, the information is rendered unreadable by all third parties.

Usability and application performance

The usability of an initial WOAP prototype (prior to the platform presented in I) was assessed through informal software usability feedback, which in the literature is sometimes dubbed Guerilla testing[56]. Guerilla testing is a rapid prototype testing method employing end-users in the intended context of use. The method can be characterized as

an inexpensive usability method feasible for non-professionals. Such usability sessions are short (~10 minutes), often informal, and thus well suited for iterative processes by enabling quick execution and analysis to feed into the next development cycle. An unspecified number of radiology consultants and residents at the Department of Radiology at Aalborg University Hospital participated in the test of the WOAP prototype. The prototype was loaded with still ultrasound images of the abdominal aorta. Participants were informed of test procedure principles, including the common phrase – "it is not you we are testing; it is the prototype," and encouraged to talk aloud about how they were experiencing the platform while using it. Test participants' interaction with the application was observed, and upon completion of the test, they were asked about their experience and any feedback/questions. When the initial version of the WOAP was deemed ready for deployment, a small feasibility study was conducted. In this study, three participants reported no problems accessing the application. They could complete the case readings without any further instructions than the one that came in the invitational e-mail and the WOAP[57]. As part of the development process, the Mulrecon imaging viewer, as specified in II was also informally tested as described above.

More formally, the usability of WOAP with accompanying imaging viewer was assessed at the conclusion of study 4. Study participants (n=15) completed the industry-standard System Usability Scale (SUS), a 10-item questionnaire to measure perceived system usability and learnability. This SUS provides a usability score that can range from 0 to 100.

This scale has demonstrated that it can be used to assess nearly any technology, so any number of devices can be evaluated with this instrument[58,59]. Furthermore, the SUS has been used in many studies, and hence its properties are well-known, with well-established benchmarks for comparative analysis[60,61]. In terms of interpreting scores, it should be emphasized that a SUS score is not simply a percentage, as research has shown that a SUS score of 68 should be considered average. To ease interpretation of SUS scores, the SUS has been mapped to a seven-point Likert scale with descriptive adjectives[59]. We used a slightly modified version of the SUS where the question "*I think that I would like to use this system frequently*" was changed to "*If I were to participate in other imaging observer agreement studies, I would like to use the DICOM viewer again.*" (Fig. 3.4). This type of change has been demonstrated to have a negligible impact on the validity or reliability of the SUS[62]. Additionally, a review of observers who had agreed to participate in either study reported in I and III-IV but who still did not complete case assessments was conducted. Moreover, based on measurement metrics from studies 2 and 3, the performance of the WOAP in terms of download speed was evaluated.

Finally, the Mulrecon DICOM viewer reported in II and subsequently enhanced with

GPU-acceleration as employed in study 2 and 3 was tested in terms of stack scroll rendering speed using imaging data from study 2. Details of this speed test are described in Appendix A.

Regarding the DICOM viewer (system) which you used to perform case assessments for each of the following 10 statements please mark the box on the 5-point likert scale that you are most in agreement with (as a refresher to the DICOM viewer, please refer to the demonstrational VIDEO):

	Strongly disagree 1	2	3	4	Strongly agree 5
1. If I were to participate in other imaging observer agreement studies, I would like to use the DICOM viewer again.	☐	☐	☐	☐	☐
2. I found the system unnecessarily complex.	☐	☐	☐	☐	☐
3. I thought the system was easy to use.	☐	☐	☐	☐	☐
4. I think that I would need the support of a technical person to be able to use this system.	☐	☐	☐	☐	☐
5. I found the various functions in this system were well integrated.	☐	☐	☐	☐	☐
6. I thought there was too much inconsistency in this system.	☐	☐	☐	☐	☐
7. I would imagine that most people would learn to use this system very quickly.	☐	☐	☐	☐	☐
8. I found the system very cumbersome to use.	☐	☐	☐	☐	☐
9. I felt very confident using the system.	☐	☐	☐	☐	☐
10. I needed to learn a lot of things before I could get going with this system.	☐	☐	☐	☐	☐

Fig. 3.4: The System Usability Scale employed in IV. Note item 10 shows, "*If I were to participate in other imaging observer agreement studies, I would like to use the DICOM viewer again.*" in place of the original, "*I think that I would like to use this system frequently.*" (Borgbjerg 2022).

Clinical application studies

Table 3.6 presents the demographic data for the study subjects (studies 1-3), and Table 3.7 summarizes the imaging protocol parameters, whereas Table 3.8 outlines the study design characteristics.

Study 1 (presented in I) primarily aimed to determine the reproducibility of ultrasound-based determination of maximum aortic diameter with the three principal methods of caliper placement concerning the aortic wall: leading to the leading edge (LTL), inner to inner edge (ITI), and outer to the outer edge (OTO) (Fig. 3.5). Ultrasound still images were used. Secondarily, the mean difference between the OTO, ITI, and LTL diameters and the impact of using either of these methods on abdominal aortic aneurysm (AAA) prevalence in a screening program was assessed.

In study 2 (presented in III), the interchangeability of ultra-low-dose non-contrast CT (ULDNC-CT) and CT for maximal abdominal aortic diameter assessment was investigated using double oblique reconstructions and a low-dose simulation technique.

Finally, study 3 (presented in IV) primarily quantified the level of intra- and interobserver variability in CT-based measurements of ductal- and gland diameters in chronic pancreatitis patients. Secondarily, sources of measurement variability were assessed.

	Study 1 (I)	Study 2 (III)	Study 3 (IV)
Number of subjects	50	50	50
Subject mean age (years, ±SD)	70, ±2.8	67.7, ±8	60.5, ±11.8
Sex (F/M)	50 M	21 F/ 29 M	15 F / 35 M
Subject BMI ((kg/m2), ± SD)	26.3 ±3.5	30, ±5.1	23.4, ±4.1
Inclusion/exclusion criteria	Men aged 65-74 years	Age over 50 years and a non-operated abdominal aorta	Diagnosis of definitive CP according to M-ANNHEIM diagnostic criteria Stents/tubes/severe organ derangement Previous pancreatic surgery

Table 3.6: Overview of study subjects' demographic data

	Study 1 (I)	Study 2 (III)	Study 3 (IV)
Modality	Ultrasound	CT	CT
Imaging system	GE Logiq E, 4 MHz curved transducer	Siemens Somatom Definition 64 slice	GE Lightspeed VCT 32 slice / GE Lightspeed VCT 64 slice
Tube voltage (kV)	N/A	120	120
Tube current time product (mAs)	N/A	Automatic tube current modulation with quality reference mAs of 220	Automatic tube current modulation 200-750 mAs
Contrast enhancement	N/A	Non-contrast /intravenous contrast timed for arterial phase	Intravenous contrast timed for portal venous phase
Reconstruction algorithms	N/A	I31F, SAFIRE 3	FBP, standard soft tissue kernel
Reconstructions	N/A	2 mm axial sections at 1.0 mm reconstruction increments	2 mm axial sections at 2.0 mm reconstruction increments

Table 3.7: A summary of imaging protocol parameters for acquisitions performed in each paper

	Study 1 (I)	Study 2 (III)	Study 3 (IV)
Study design	Retrospective	Retrospective	Retrospective
Total number of two-point caliper assessments	5400	1400	1600x4 diameters = 6400
Number of assessments	150 (interobserver) / 300 (intraobserver)	200	100x4 = 400
Type of assessment	Diameter measurement of the infrarenal abdominal aorta	Diameter measurement of the abdominal aorta	Four largest diameters: PDhead, PDbody MPDhead, and MPDbody
Imaging plane	Axial perpendicular to the flow direction of the aorta	Axial double oblique reconstruction perpendicular to the flow direction of the aorta	Axial only
Measurement technique	Maximum diameter using the OTO, ITI, and LTL technique	Maximum diameter using the OTO technique	Diameter measurements preferably perpendicular to the center axis of the pancreas
Number of participating institutions/countries	5 Danish hospitals / 1	1 Danish hospital / 1	11 Hospitals / 10
Number of observers	18 interobserver / 12 intraobserver	7	16
Observer experience (years, ±SD, range)	11.8, ±9, 4-37 (Ultrasound)	7.9, ±3.2, 4-15 (CT)	7.9, ±3.2, 4-15 (CT)
Observer profession/employment	Radiology consultants (n=11) Radiology residents (n=7)	Radiology consultants (n=4) Radiology residents (n=3)	Radiology consultants (n=10) Radiology residents (n=2) Clinical pancreatologists (n=4)
Observer eligibility	> 3 years of radiological experience Perform ultrasound at least once a month during the previous 12 months	> 3 years of radiological experience and experience with double oblique MPR. Having interpreted CTs depicting the abdominal aorta at least once in the last 12 months.	Experience in reading pancreatic CT images
Coaching session	Web-based written instruction	Web-based written instruction and demonstration videos	Web-based written instruction and demonstration videos

Table 3.8: Study design characteristics, observers, and measurement comparisons for the individual clinical application studies

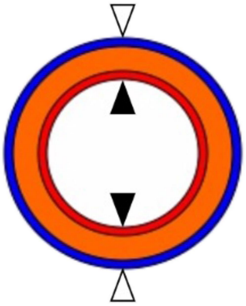


Fig. 3.5: Schematic transverse image of the abdominal aorta. The inner red circle represents the tunica intima, the orange area represents the tunica media, and the outer blue circle represents the tunica adventitia. The three principal methods of caliper placement in ultrasound assessment of maximum abdominal aortic diameter are inner to inner edge (ITI, solid black arrowheads), leading to the leading edge (LTL, downward black hollow arrowhead to downward solid arrowhead, and outer to the outer edge (OTO, hollow black arrowheads).
Reprinted from European Journal of Vascular and Endovascular Surgery 2018[48], copyright, by permission of Elsevier.

Patient populations and imaging studies

Patients included in study 1 came from the Viborg Vascular (VIVA) screening trial, where men aged 64-74 were screened for an AAA using ultrasound. Participants (n=50) were randomly selected from the VIVA screening database. One ultrasound image of the infrarenal aorta of each patient was included in the study. The database from the VIVA screening trial is approved for research purposes by the regional scientific ethics committee and by the Danish Data Protection Agency.

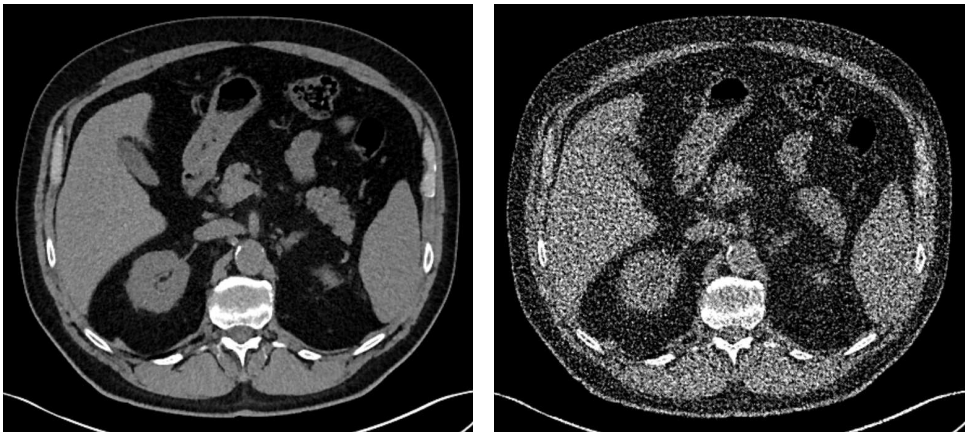
In study 2, 50 consecutive patients who underwent a CT as part of the workup for suspected renal artery stenosis at Aarhus University Hospital were identified in the PACS system. Non-contrast and arterial phase series were obtained for all patients using the same CT system (Table 3.7). The study did not access patients' electronic health records, the departmental research committee approved the study, and the regional ethics committee waived the need for informed patient consent.

Patients in study 3 came from The Scandinavian Baltic Pancreatic Club (SBPC) Database (http://sbpcforumofexcellence.com). All 50 patients included had contrast-enhanced abdominal CT scans performed at Aalborg University Hospital using two CT systems

(Table 3.7). Patients were included with approval by the Danish Data Protection Agency. All imaging studies had identifiable subject information removed.Observers were recruited by e-mail invitation, and all gave written informed consent to participate. Details of the participating observers for the three application studies are provided in table 8. No compensation was provided to the observers.

CT low dose simulations

To generate ULDNC-CT datasets for study 2, a previously demonstrated simulation technique that relies on normal dose CT images without needing raw sinogram data was used (Fig. 3.6)[30]. In this approach, noise samples were obtained from scanning a 320 mm diameter polymethylmethacrylate phantom filled with water with attenuation equal to soft tissue. A tube current of 17 mAs was used, which is the lowest possible setting in the Siemens Somatom Definition AS 64 system. Images were reconstructed with filtered back projection and a soft tissue kernel. The noise data was subsequently introduced into the non-contrast clinical patient scans to approximate ultra-low-dose CT reconstructed with filtered back-projection.



A B

Fig. 3.6: The low-dose simulation technique used in study 2. Axial slices from the original normal-dose non-contrast CT A) and simulated low-dose non-contrast CT B) at the origin of the right renal artery. The transition between the abdominal aorta and the left crus of the diaphragm is substantially less conspicuous on the simulated low-dose non-contrast CT owing to the increased noise. Adapted from III.

Case assessments using the WOAP

The WOAP was used in all three clinical application studies.
In study 1 (presented in I), two measurement sessions were conducted. In the first round, each observer (n=18) performed measurements of the maximum diameter of the 50 ultrasound still images in random order with the OTO, ITI, and LTL techniques. Hence, a total of 150 measurements for each observer gave 2700 first session measurements and 900 caliper placements for each of the three methods to measure the aortic diameter. In the second session, randomized measurements of the 50 images were repeated in a subset of observers (n=12) with the OTO, ITI, and LTL methods for a total of 1800 measurements and 600 caliper placements for each of the three methods to measure the aortic diameter. No image manipulation tools were made available during the case assessment.

In study 2 (presented in III), observers performed measurements of the maximum abdominal aortic diameter in any direction using MPR in an imaging plane perpendicular to the aortic centerline as recommended in guidelines. Commonly available DICOM viewer functionality was enabled (Table 3.7).
Only the aorta below the coeliac trunk to the aortic bifurcation was evaluated. Diameters were measured from the outer to the outer wall (OTO) of the aorta. In the case that the abdominal aorta diameter was less than 2.5 mm (i.e., no aortic ectasia/AAA), observers were instructed to place calipers in the maximum cross-section perpendicular to the aortic centerline just below the coeliac trunk. A representative image of maximum abdominal aortic diameter (Dmax) caliper placement using the web-based DICOM viewer is shown in Fig. 3.7. Two measurement sessions were conducted. In the first session, each observer (n=7) performed measurements of the maximum aortic diameter of the 50 ULDNC-CT and 50 CT angiography (CTA) scans for a total of 100 measurements each. In the second session, measurements of the 50 ULDNC-CT and 50 CTA scans were repeated for an additional 100 measurements completed by each observer. In each reading session, the review of the 50 simulated ULDNC-CT datasets preceded the assessment of the 50 corresponding CTA datasets. The x-y coordinates of each caliper placement were saved in the database.

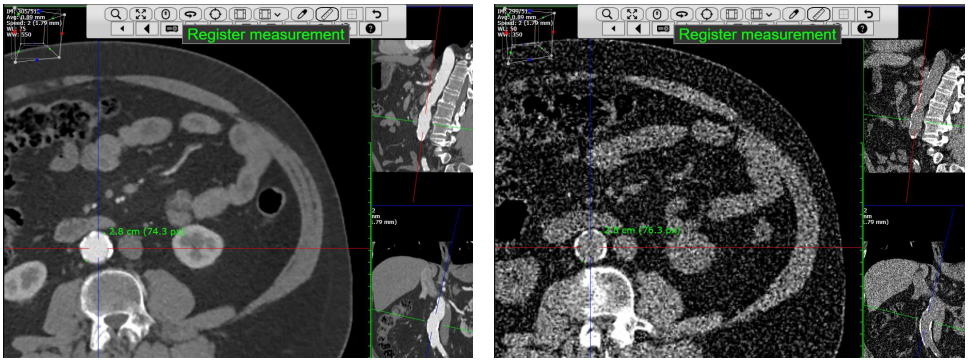


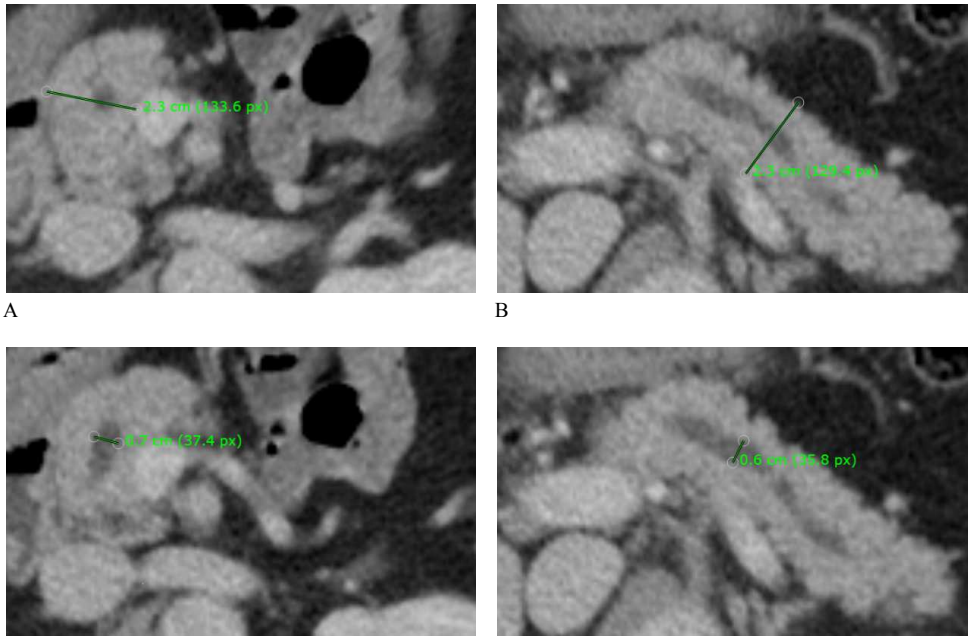
Fig. 3.7: The web-based DICOM viewer used in study 2 (III). Measurement of maximum diameter of the abdominal aorta using double-oblique reconstruction and the centerline technique. A) CT angiography and B) ultra-low-dose non-contrast CT. Adapted from III.

A key image of each caliper placement was automatically saved in the platform database. Measurements that deviated more than 3 mm from a given abdominal aortic mean diameter were investigated to identify sources of inaccuracy. The time consumption to obtain ULDNC-CT and CTA diameter measurements, respectively, was also registered.

In study 3 (presented in IV), pancreatic ductal- and gland diameter measurements were performed on axial images only and according to the instructions from the SBPC imaging module, which is described in Lisitskaya et al.[63] and at http://sbpcforumofexcellence.com. Four diameters were assessed where measurements should preferably be perpendicular to the center axis of the pancreas (Fig. 3.8); largest diameter of the pancreatic head (PDhead), the largest diameter of the pancreatic body (PDbody), largest main pancreatic duct diameter in the pancreatic head (MPDhead), and largest main pancreatic duct diameter in the pancreatic body (MPDbody). The default three orthogonal image stacks were presented to observers, and the ability to perform oblique reconstructions was disabled. Otherwise, commonly available DICOM viewer functionality was enabled. Two measurement sessions were conducted. In the first session, each observer measured the four diameters of the 50 CT scans in randomized order, in total 200 measurements each. Each observer repeated all 200 measurements in a randomized order in the second session. The two reading sessions were separated by a minimum of two weeks. The reading time of each case was registered from when a scan had loaded until all measurements were completed.

Furthermore, a key image of each caliper placement was automatically saved in the

platform database, and the x-y-z coordinates of each two-point caliper placement were registered.



A B C D

Fig. 3.8: Pictorial presentation of two-point caliper placement in an axial CT scan (study 3 - IV) of the pancreas with intravenous contrast in the portal venous phase for (A) PDhead, (B) PDbody, (C) MPDhead, and (D) MPDbody. Adapted from IV.
Abbreviation: PDhead = diameter of the pancreatic head; PDbody = diameter of the pancreatic body; MPDhead = main pancreatic duct diameter of the pancreatic head; MPDbody = main pancreatic duct diameter of the pancreatic body.

Statistical analysis

The following section primarily outlines the statistics used in the thesis papers in relation to assessing observer agreement of continuous in a multi-observer setup. More statistical details can be found in papers I, III-IV.

Generally, summaries of continuous variables are represented by means and standard deviations (SDs), and paired t-tests were used to assess the difference in mean diameters and case reading time.

The Bland- Altman approach is the most commonly used method for assessing agreement on the measurement of a continuous variable[64]. This method plots differences between two observers (or an observer performing repeated measurements, i.e., intraobserver agreement) against respective means together with 95% limits of agreement (LoAs). The LoA estimates the range over which one would expect 95% of differences to lie. The approach can be used to access both intra- and interobserver agreement, and the width of the LoA is generally likely to be larger in the latter case. As previously mentioned, a major limitation of the classical Bland–Altman plot is that it only applies to a situation with two observers or methods.

Intra-class correlation coefficient (ICC) is another commonly utilized statistical method for evaluating measurement methods by providing a reliability index for evaluating continuous variables that reflects both the degree of correlation and agreement between measurements[65]. In contrast to the Bland-Altman method, the ICC can accommodate multiple observers. Nevertheless, it is not an ideal method for evaluating observer variability; the ICC reveals little about the degree of discrepancy nor supplies information to investigate whether the variability may change with the magnitude of measurements (e.g., to reveal that the diameter of large abdominal aneurysms is less precisely estimated compared to smaller ones). Further, deciding what value constitutes sufficiently high reliability is often made subjectively. Lastly, since the ICC is heavily dependent on between-subject variation, it may produce a high value simply due to a heterogeneous patient group.

To extend Bland-Altman's method to provide a simple statistical approach for evaluating the agreement of continuous variables between multiple observers, Jones et al. suggested limits of agreement with the mean (LOAM)[22,64]. For data visualization, Jones et al.

proposed an agreement plot of the observed differences against the observed subject-specific average, and this plot is equipped with horizontal lines representing the 95% LOAM. However, the LOAM by Jones et al. (henceforth Jones LOAM) does not include any possible variation due to measurements made by different observers. Hence, the Jones LOAM can be interpreted as *how much a given observer's measurement may plausibly deviate from the mean of all measurements performed by that particular observer on the specific subject*.

We reformulated the Jones LOAM under an additive two-way random effects model described in a paper by Christensen et al. [66]. In this model, the total variation present in a set of measurements by different observers is partitioned into components attributable to different sources of variation: the inter-subject (σ_A, i.e., variation between the true values for subjects), inter-observer (σ_B, i.e., varying bias between observers where observers consistently measure high or low), and residual variance (σ_E). It is assumed that the effects of these variance components are added, and regarding the residual variance component, this can also be termed the *error variance* and is related to measurement repeatability. This reformulation allows one to consider multiple measurements per observer and define confidence intervals (CIs) for the 95% LOAM and the individual variance components as recommended in the litterature[67]. An example agreement plot from I is shown in Fig. 3.9.

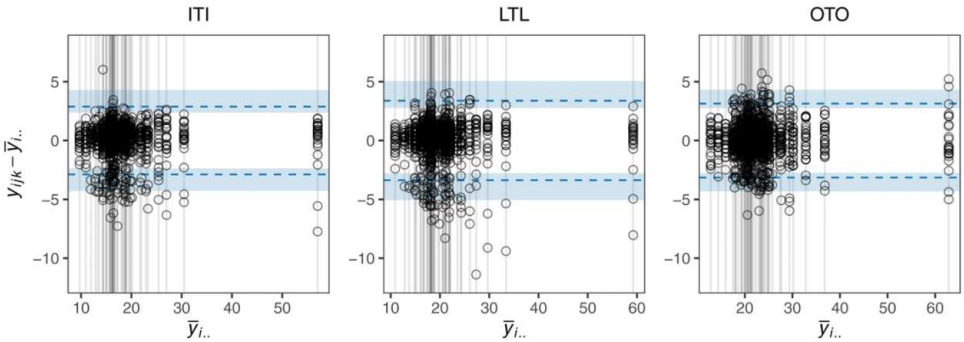


Fig. 3.9: Illustration of the LOAM approach based on data reported in I. Agreement plots for each of the three methods (OTO, LTL, and ITI) used to measure the aortic diameter along with the estimate (dashed line) and the 95% CI for the 95% LOAM (shading). Observed differences of repeated measurements of the 12 observers $d_{ij}=y_{ij}-\underline{y}_{i\cdot\cdot}$ is plotted against the subject-specific average $\underline{y}_{i\cdot\cdot}$ across observers.
Abbreviation: OTO = outer to outer; LTL = leading to leading edge; ITI = inner to inner.
Reprinted from BMC Medical Research Methodology 2020[49,66], copyright, by permission of Springer.

In III-IV, LOAMs are calculated using the reformulated approach by Christensen et al. and equipped with an asymmetric but approximate CI; see Christensen et al. for calculation and discussion on the quality of such CIs. The reformulated 95% LOAM represents how much a given observer's measurement may plausibly deviate from the mean of all observers' measurements on the specific subject (i.e., a measure of reproducibility as the intra- and interobserver variation is combined). As part of an agreement analysis, the order of magnitude of the variance components can be compared to elucidate the main sources of disagreement, e.g., is the magnitude of the inter-observer variance component minor relative to the residual variance to such a degree that different observers can safely perform the specific measurements? Furthermore, it should be emphasized that obtaining multiple measurements per observer and, in particular, increasing the number of observers allows more precise estimates of LOAM and associated variance components.

In I, we used the original Jones LOAM for assessing agreement. The paper was accepted in the European Journal of Vascular and Endovascular Surgery[48]. Later, as we reformulated the LOAM in the paper by Christensen et al., a corrigendum was submitted to the European Journal of Vascular and Endovascular Surgery (Appendix B). We presented results based on the reformulated LOAM. Thus, the results presented in this thesis are based on this corrigendum.

In III and IV, for assessing agreement, we used reformulated LOAMs calculated based on first and second session measurements and equipped with an asymmetric and approximate CI. Based on a fixed number of subjects/patients and initial estimates of the intra- and inter-observer variation, Christensen et al. provided a formula for determining how many observers are needed to obtain an expected width of the 95% CI for the 95% LOAM[68]. This formula was used in III.

For papers III and IV, to enable easier comparison with prior studies, we also calculated Bland-Altman LoAs. When evaluating agreement in a multi-observer setup, it is possible to present multiple Bland–Altman plots for each pairwise comparison of observers, but this becomes difficult to present and interpret for more than four raters. Instead, we used an approach where multiple LoAs are calculated based on possible observer pairs as utilized in a paper by Kakinuma et al.[69]. Using this approach, Bland-Altman LoAs for intra-observer pairs of repeated diameter measurements by each observer were calculated. In addition, LoAs of first session diameter measurements between all possible inter-observer pairs were evaluated.

Moreover, a web-based statistical module written in JavaScript for IV was developed and coupled with the WOAP. The module allows interactive exploration of measurements in an agreement plot based on the LOAM method. Measurement points on the agreement plot are coupled to the corresponding key image saved in the database enabling a swift display of the caliper placement. In addition, the module provides links to the individual CT studies where calipers placed by the 16 observers are shown. Hence, this module enables exploration of measurement outliers and, compared to a static plot, can more clearly delineate measurements by a given observer than all other measurements (Fig. 3.10).

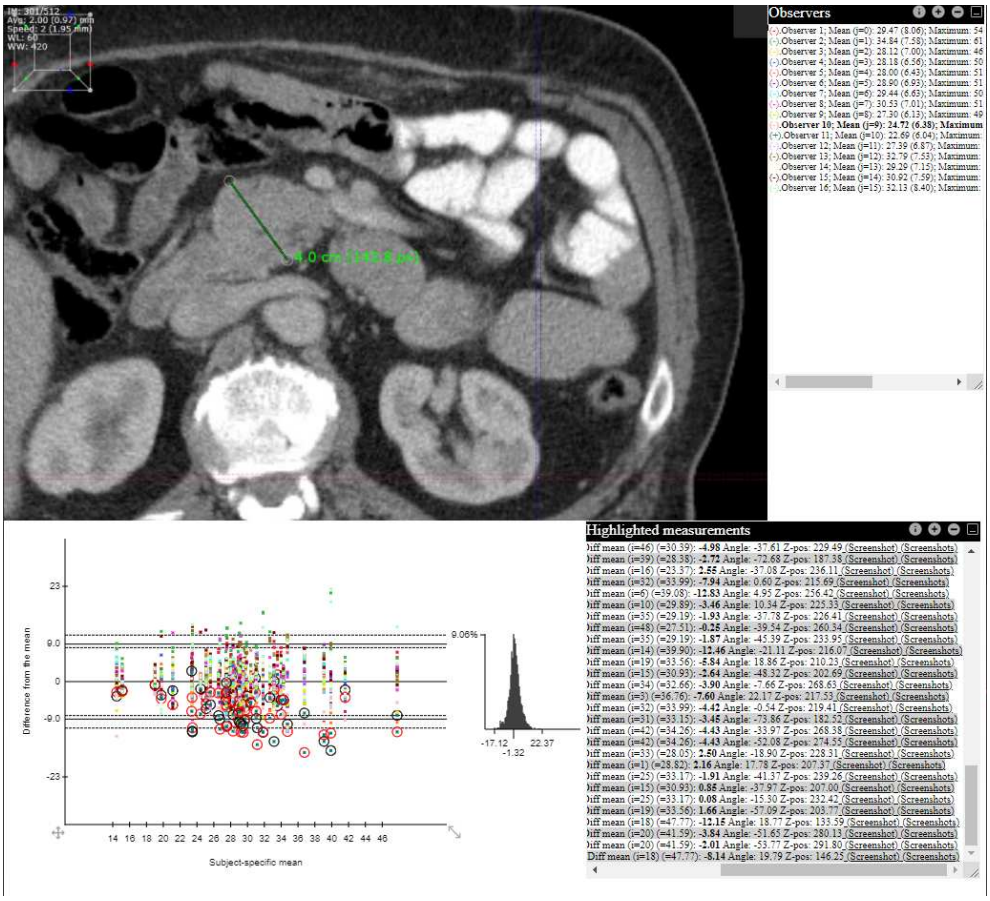


Fig. 3.10. The interactive web-based statistical module with an agreement plot from study 3. Measurements of PDhead of 16 observers are shown. The measurements by observer 10 have been highlighted in the plot, and a key image of the measurement on CT case 4 by that observer is displayed (Borgbjerg 2022). Abbreviation: PDhead = diameter of the pancreatic head.

Results

In this chapter, the thesis results are summarized in relation to the aims. More detailed results are presented in I-IV.

AIM I

Aim: To develop and improve a web-based database application for observer authentication, study management, and storage of evaluation results (I, III, IV)

<u>Key results:</u>

- An easily accessible platform-independent web-based application for facilitating observer performance studies was developed
- Three clinical application studies were completed which demonstrated:
    - Gradual improvements were implemented from display of 2D images only to accommodation of volumetric datasets.
    - Application can be adapted to the specific needs required for answering different study questions in observer agreement studies.
    - Extensive data can be collected for subsequent analysis in an interactive manner.

Interpretation:

It was possible to develop a web-based platform for facilitating observer performance studies that allowed invited observers to be scattered in localization where case readings can be separated in time and not necessarily have to be completed all at once. Using the web-based database, automated extensive data collection concerning case readings was possible.

AIM II

Aim: To determine reproducibility using the web-based application in the ultrasonographic assessment of maximum abdominal aortic diameter with three principal methods of caliper placement (I)

<u>Key results:</u>

- Eighteen observers each completed one session of 150 measurements for a total of 2700 caliper placements. Twelve observers each completed a second measurement session of 150 measurements for a total of 1800 caliper placements.

- The mean OTO aortic diameter was 23 mm (95% CI 21-25 mm), LTL diameter 20 mm (95% CI 18.5-22.3 mm), and ITI diameter 18 mm (95% CI 16.1-19.9 mm).
- OTO demonstrated 95% LOAM, σ_A, σ_B, and σ_E of 3.2 (2.8, 4.3), 7.2 (5.7, 8.6), 1.1 (0.7, 1.6), and 1.2 (1.2, 1.3) mm (Fig. 3.9).
- LTL demonstrated 95% LOAM, σ_A, σ_B, and σ_E of 3.4 (2.8, 5.1), 6.9 (5.5, 8.3), 1.5 (0.8, 2.1), and 1.0 (1.0, 1.1) mm (Fig. 3.9).
- ITI demonstrated 95% LOAM, σ_A, σ_B, and σ_E of 2.9 (2.4, 4.3), 6.8 (5.4, 8.1), 1.2 (0.7, 1.8), and 0.9 (0.9, 0.9) mm (Fig. 3.9).
- Mean differences were: 5.0 mm (95% CI 2.3–7.8, $p < .05$) between OTO and ITI measurements, 2.6 mm (95% CI –0.2-5.4, $p < .05$) between OTO and LTL measurements, and 2.4 mm (95% CI −0.3-5,1, $p < .05$) between LTL and ITI measurements.
- Mean difference estimations regarding LTL-ITI and OTO-ITI applied to all 18,698 individual ITI measurements in the VIVA AAA screening trial demonstrated that 756 (4.0%) and 1110 (5.9%) AAAs would have been diagnosed, respectively, if LTL or OTO had been used instead of ITI (615, 3.3%). Almost one-fifth of abdominal aortas would be considered ectatic if the OTO method was used, compared with 2.6% with the ITI method.

Interpretation:
The ITI method in assessing maximal aortic diameter demonstrated superior reproducibility compared to OTO and LTL, whereas ITI and LTL demonstrated the lowest residual variance, which approximates better repeatability. The choice of caliper placement method affects the prevalence of AAAs in screening programs.

AIM III
Aim: To develop an easily accessible web-based DICOM viewer for interactive visualization of volumetric datasets (II)

<u>Key results:</u>
- The Mulrecon viewer has an interface and functionality akin to a PACS.
- The developed viewer was platform-independent and compatible with all major internet browsers
- The viewer has support for the DICOM format.
- The viewer implements hardware-accelerated rendering.
- Rendering speed: please see aim VI.

- The viewer implements functionality consistent with display requirements defined by the Royal College of Radiologists (Appendix C).

Interpretation:
It was possible to develop a web-based DICOM viewer for visualization of volumetric DICOM datasets with functionality and real-time rendering speed comparable to desktop-based PACS systems.

AIM IV
Aim: To assess the interchangeability of ultra-low-dose non-contrast CT and standard-dose CT angiography using the improved web-based application in terms of accuracy and reproducibility in determining maximum abdominal aortic diameter. (III)

<u>Key results:</u>

- Seven observers each completed two sessions of 100 measurements (corresponding to 50 ULDNC-CT and 50 CTA), each using an MPR technique for a total of 1400 caliper placements.
- The mean diameter was 24.0 (±0.4, 17.6-37.6) mm for CTA and 25.0 (±0.5, 18.7-37.4) mm for ULDNC-CT.
- A significant mean difference of 1.0 mm (95% CI 0.8–1.2, $p < 0.001$) between ULDNC-CT and CTA was found.
- ULDNC-CT demonstrated 95% LOAM, σ_A, σ_B, and σ_E of 2.3 (2.2-3.1), 3.5 (2.8-4.3), 0.6 (0.2-0.9), and 1.1 (1.1-1.2) mm (please see Fig. 3B from III).
- CTA demonstrated 95% LOAM, σ_A, σ_B, and σ_E of 2.3 (2.1-3.5), 3.9 (3.1-4.7), 0.7 (0.3-1.1), and 1.0 (1.0-1.1) mm (please see Fig. 3A from III)
- The average time to obtain a maximum abdominal aortic diameter measurement was 80 seconds (95% CI 43–119) for CTA and 112 seconds for ULDNC-CT (95% CI 75–150), yielding a mean difference of 32 seconds (95% CI 19–44, $p < 0.001$).
- The Bland-Altman LoA intra- and interobserver pairs that went beyond the clinically acceptable range of +/- 5 mm only did so with a small margin.

Interpretation:
Non-contrast CT scans at ultra-low-dose levels are interchangeable with gold-standard CTA to assess abdominal aortic diameter. Measurements can be completed in a timely fashion compatible with clinical practice.

AIM V

Aim: To quantify the level of observer variability using the improved web-based application in CT-based measurements of ductal- and gland diameters in chronic pancreatitis (IV)

Key results:

- Sixteen observers each completed two sessions of 200 measurements (corresponding to 50 CTs measuring PDhead, PDbody, MPDhead, and MPDbody) each for a total of 6400 caliper placements.
- PDhead demonstrated 95% LOAM, σ_A, σ_B, and σ_E of 9.1 (8.2 to 11.2), 6.0 (4.8 to 7.3), 2.9 (1.9 to 4.0), and 3.7 (3.6 to 3.8) mm (please see Fig. 3A from IV)
- PDbody demonstrated 95% LOAM, σ_A, σ_B, and σ_E of 5.1 (4.9 to 5.8), 4.0 (3.2 to 4.9), 1.1 (0.7 to 1.6), and 2.4 (2.3 to 2.5) mm (please see Fig. 3B from IV)
- MPDhead demonstrated 95% LOAM, σ_A, σ_B, and σ_E of 3.2 (3.1 to 3.4), 2.3 (1.9 to 2.8), 0.4 (0.2 to 0.6), and 1.6 (1.6 to 1.7) mm (please see Fig. 3C from IV)
- MPDbody demonstrated 95% LOAM, σ_A, σ_B, and σ_E of 2.6 (2.5 to 2.9), 2.8 (2.2 to 3.3), 0.5 (0.3 to 0.7), and 1.3 (1.2 to 1.3) mm (please see Fig. 3D from IV)
- Studying key images revealed that major sources of measurement variation were failure to locate and measure at the widest level of the pancreatic head/body and failure to measure perpendicular to the axis of the pancreatic head/body.
- A high and moderate correlation between measurement variation and mean angle difference for PDhead and PDbody, respectively, was seen.
- A moderate correlation between measurement variation and mean midpoint distance for PDhead and PDbody was seen.

Interpretation:

Two-point pancreatic measurements are subject to substantial intra- and interobserver variability among specialists. The findings question the implementation of two-point measurements as the basis for imaging scoring systems in chronic pancreatitis.

AIM VI

Aim: To assess the usability and performance of the developed platform (I, III, and IV)

Key results:

- Based on 15 participants from IV, the average System Usability Scale score (SD, range) of the WOAP was 84 (15, 53-100).

- All participants (n=18) who had agreed to participate in I completed case assessments.
- Two out of 9 participants who had agreed to participate in study 2 failed to complete case assessments. The stated reason for withdrawal in both cases was that the participants found that they could not find the necessary time to complete case assessments after further consideration.
- One out of 17 participants who had agreed to participate in study 3 failed to complete case assessments. The stated reason for withdrawal was that the participant could not access the WOAP through the hospital firewall. The participant was not interested in trying to resolve this technical issue.
- CT studies of III had a mean (SD, range) number of images and a total study size (SD, range) of 346 (153.3, 101-639) and 182 megabytes (81, 53-337), respectively.
- Based on 7 participants in III, the average time (SD, range) to download a study in III was 28.2 (31, 4.9-277) seconds, yielding an average download speed of 6.5 megabyte/second.
- CT studies of IV had a mean (SD, range) number of images and a total study size (SD, range) of 149 (25.3, 100-250) and 78.7 megabytes (13.4, 52.8-132.1), respectively.
- Based on 16 participants in IV, the average time (SD, range) to download a study in IV was 20.9 (19.9, 5.1-191.2) seconds, yielding an average download speed of 3.8 megabytes/second.
- Based on a study 2 CT dataset, the GPU-accelerated DICOM viewer demonstrated a stack scroll speed of 28 frames per second (Appendix A).

Interpretation:
The download and rendering speed of the WOAP was satisfactory for the completion of observer agreement studies. The perceived usability of the WOAP and Mulrecon DICOM viewer was in the range between good (SUS=71.4) and excellent (SUS=85.5) according to the SUS adjective rating and compares favorably to well-known, highly used products such as Internet Browsers (SUS=81.1) and microwave ovens (SUS=87.2).

Discussion

This thesis presents a web-based application to facilitate observer performance studies in imaging research and development. In this discussion section, the main results in the present thesis will be combined, with a focus on both the developmental aspects and the application in clinically relevant observer performance studies and with a comparison with findings in the literature. Subsequently, the implications of the research results will be explored, followed by an outline of limitations. The chapter concludes with a discussion of future perspectives, especially about the adaptation of the web-based observer performance platform in clinical imaging studies. For more disease-specific details and associated clinical perspectives concerning abdominal aortic aneurysms and imaging biomarkers in pancreatitis, we refer the reader to I, III, and IV.

In I, we assessed the reproducibility of the three principal caliper placement methods in determining maximum abdominal aortic diameter using ultrasound. We found superior reproducibility of the ITI method and recommended a continuation of current screening programs as well as an adaptation by imaging departments to the ITI method. In III, we sought to determine whether ULDNC-CT can be used instead of the gold-standard CT angiography for the assessment of maximal abdominal aortic diameter. We concluded that ultra-low-dose non-contrast CT exhibited similar accuracy and reproducibility of measurements compared with CTA for assessing maximal abdominal aortic diameter, supporting that ULDNC-CT can be used interchangeably with CTA in the lower range of aortic sizes. In IV, we quantified the level of intra- and interobserver variability in CT-based measurements of ductal- and gland diameters in chronic pancreatitis patients. We demonstrated substantial intra- and interobserver variability in two-point measurements, even among specialists. We concluded that our findings question the implementation of two-point measurements as the basis for imaging scoring systems in chronic pancreatitis. The DICOM viewer developed as part of II proved viable for conducting observer performance studies, evidenced by the implemented functionality, high completion rate of case assessments, and excellent System Usability Scale score. Consequently, studies can be performed in concordance with display requirements defined by the Royal College of Radiologists.

As outlined in the introductory literature review, it is well-recognized that observer agreement studies involving few observers may not expose the true extent of variability between observers who interpret imaging studies in clinical practice. This continues to be

an open problem in the assessment and implementation of new imaging techniques; for example, a recent review of techniques for CT dose reduction emphasized that even though multi-observer studies best demonstrate radiologists' performance, they are resource-intensive and impractical to conduct by traditional means[70]. In this context, we believe that the results presented in IV on two-point pancreatic CT measurements adeptly illustrate the necessity of performing multi-observer studies when evaluating agreement. Here it is essential to keep the limitations of the Bland-Altman method in mind, which is the most commonly deployed statistical method when considering agreement in continuous variables; it only applies to a situation with two observers or methods. Consequently, the limits of agreement pertain to the two specific observers having performed measurements, and thus results can not readily be extrapolated to the whole population of potential readers. While a number of the interobserver pairs reported in IV are seen to have an excellent agreement, the overall measurement variability between observers was substantial, and several interobserver pairs demonstrated surprisingly large variation. Hence, if only two randomly selected observers had been included in the study, the likelihood of concluding differently as to the clinical utility of two-point measurements was considerable.

The III study also included a relatively large number of observers but illustrates a different point when evaluating observer agreement. The ULDNC-CT and the CTA exhibited similar LOAM, and in addition, LoAs of the Bland-Altman interobserver pairs were also similar, and only a few were outside the interval of what has been deemed clinically acceptable difference. We believe this study is an example of how one can use a web-based platform to reasonably establish the interchangeability of imaging modalities according to recommendations in the literature.

To put things into perspective, in terms of research practice, only a few prior observer agreement studies have included many observers from different institutions comparable to the studies in I and IV, with the participation of 18 and 16 observers, respectively. One such example is the well-recognized single-center study by McErlean et al., which assessed the variability of CT measurements of cancer lesions in 17 observers with varying experience levels using a routine clinical PACS[71].

Emergent proposals for web-based facilitation of observer performance studies

Since the inception of this Ph.D. project, a number of web-based proposals potentially capable of facilitating observer performance studies have been described in the literature (overview in Table 5.1). Yang et al. and Rubin et al. have presented solutions where a Java-based DICOM viewer is coupled with a DICOM server[72,73]. Both solutions are intended for the development and validation of imaging biomarkers as part of oncology trials. The solutions are not purely web-based since they rely on the Java plugin, which imposes restrictions as previously described. In contrast to Yang et al., the Epad platform by Rubin et al. is available for download and can be customized by writing Java-based plugins. In a feasibility study, Hostetter et al. have demonstrated how a pure web-based DICOM viewer can be coupled to a commercial web server and, by means of web forms, allow the creation of a variety of question types[74]. However, while caliper measurements can be performed in the DICOM viewer, the registered diameters must be manually entered by observers in a free text field. Additionally, using a commercial DICOM server makes the solution inaccessible to customization by third parties. Ziegler et al. presented the OHIF viewer platform, which features a pure web-based DICOM viewer[75]. This viewer has been coupled with an open-source DICOM viewer as part of a project to annotate publicly available DICOM datasets contained in the Cancer Imaging Archive. Participants performed bidirectional measurements on cancer lesions, and data was stored in a custom database. In all, the above solutions, which are amenable to customization, are not specifically designed for observer performance studies. In order to be used for specific observer performance studies they will need complex and extensive software configuration as recognized by Ziegler et al. Hence, solutions cannot be expected to be adopted by researchers without comprehensive assistance by a capable programmer/IT professional to set up the DICOM image archive, participant authentication system, and storage of case assessment results. More recently, in 2022, Genske and Janke presented an open-source platform for performing observer performance studies in imaging[76]. It features a somewhat rudimentary DICOM user interface lacking features such as window/level adjustment, scrolling through multi-stack images, and more advanced volumetric image manipulation. In contrast to the above solutions and the WOAP, the platform does, however, implement a web content management system with a graphical user interface. This management system allows the basic setup of different observer performance studies, such as those involving multiple-alternative forced-choice and

location receiver operating characteristics methods. Several research projects have been completed using the platform, and radiologists evaluating the platform usability based on the SUS gave it a score of 83 (A rating). However, it remains to be seen how easy it will be for third-party researchers to utilize this platform on their own[77].

In comparison to the above proposals, our developed web-based platform differs in several ways. First, in contrast to the first three referenced solutions, the WOAP has been specifically designed to conduct observer performance studies. Secondly, the codebase of the back-end solution of the WOAP is comparatively much smaller, allowing more flexibility, ease, and speed of customization. This is reflected in the relatively easy adaptation of the platform setup needed to transition from application study 2 to the execution of study 3. Third, the upload and hosting of DICOM files do not require a dedicated DICOM server but merely require a standard web server.

Platform	Intended usage	Technical details	Description
Weasis-based[72]	Cancer response-assessment system to foster the development and validation of new quantitative imaging biomarkers.	Java-based plugin Implements multiplanar reconstruction	Complex and extensive software configuration Closed-source code DICOM server Orthogonal MPR
Pacsbin[74]	Provides a research platform for multi-reader multi-case studies and other imaging research	Pure web-based viewer Image stacks, but no multiplanar reconstruction	Back-end is based on commercial solution Closed-source code DICOM server No MPR
ePAD[73]	A platform for medical image annotations and quantitative analysis primarily in cancer research.	Java-based plugin Implements multiplanar reconstruction	Complex and extensive software configuration Closed-source code Customization through Java-based plugins DICOM server Orthogonal MPR
OHIF viewer[75]	To develop purpose-built applications for small subsets of patients, experiment with new imaging tools, or produce training modules.	Pure web-based viewer Implements multiplanar reconstruction	Complex and extensive software configuration DICOM server Double oblique MPR
Human Observer Net[76]	To develop a user-friendly software platform that enables efficient human observer studies in medical imaging with flexibility of study design.	Web-based viewer that requires a web server supporting docker images.	Web content management system. No MPR

Table 5.1: Web-based proposals potentially capable of facilitating observer performance studies.

Implications for use in clinical research studies

The results obtained as part of developing and applying the developed WOAP suggest that the platform is a viable solution for conducting observer performance in imaging research and development. By switching from observer performance studies on a PACS or stand-alone DICOM viewer to the WOAP, a virtually unlimited number of observers and cases can be included in a given study. The proposed solution can help decrease the time

needed of both researchers and participating observers. Given the comparatively light codebase and the ability to run the WOAP on a standard web server, researchers can relatively efficiently adapt the platform to execute different studies with observers from different institutions/countries and in relatively large numbers, which is rarely reported in the literature (i.e., > 10 observers). Hence, it becomes more feasible to complete agreement studies according to recommendations such that the untapped potential of dose-saving imaging alternatives for specific diagnostic tasks can better be realized[34,78,79]. A recent example of such a task is evaluating whether a low-dose CT protocol or abbreviated non-contrast MRI can be used instead of a standard-dose CT scan in active surveillance of small renal masses[80,81,82]. The three clinical application studies included in this Ph.D. thesis all evaluated observer agreement of continuous variables; however, the WOAP is equally capable of handling other types of observer performance studies, such as those involving categorical data as well as diagnostic accuracy studies. A recent paper has re-emphasized the need to quantify observer variability in imaging endpoints of cancer trials[83]. The paper also underlines the additional need to identify reasons for observer variability, particularly those arising in radiological response assessment of studies involving immunotherapies that are quickly becoming mainstream[83]. Using traditional platforms for executing observer performance studies, these data are difficult to obtain and time-consuming to analyze. As demonstrated in papers III and IV, the ability to automatically gather extensive quantitative data from case assessments and provide an interface for identifying outliers and comparing key images can help in this regard. From a broader scientific perspective, the execution of performance studies with a greater number of observers can be an important tool seen in the light of the increasing risk of over-interpretation and "spin" in imaging research where too far-reaching conclusions are stated based on sparse data[21]. In terms of current and future radiological practice, one might ask if evaluation of human observer performance matters in an era with artificial intelligence (AI) on the rise? A 2020 survey on AI completed by members of the American College of Radiology indicated a modest penetration of AI in clinical practice with concerns regarding inconsistent performance and whether incorporation of AI will decrease productivity[84]. That radiology as a whole will be impacted by AI is certainly beyond doubt. However, best-informed opinions expect AI to function as a "co-pilot" in reducing error and repetitive tasks and not as a replacement for radiologists[85]. Thus, imaging interpretation will, at least for the foreseeable future, continue to rely on human expertise; and reader variability will remain an unavoidable reality. Currently, a plethora of narrowly focused applications of AI has appeared. It remains to be seen how success in a research setting will translate into routine clinical practice across many institutions. Radiology peers emphasize that radiologists should embrace the opportunity to guide the

development, education, regulation, and deployment of AI into the clinical arena. It is clear that the incorporation and utilization of all AI models in PACS systems are not feasible for all possible imaging evaluations. Hence, there is a need to clarify which radiological tasks with great advantages can be delegated to an AI system. The IV study reported observer agreement of two-point measurements in chronic pancreatitis patients and found substantial measurement variation even though such measurements, in principle, are fast and easy to obtain. Thus, we believe that IV is an excellent example of how the WOAP can help elucidate domains within diagnostic imaging where there is an urgent need to investigate and develop AI-based applications such as automatic pancreatic gland volume segmentation.

Limitations

The conundrum that utilization of the web-based solutions listed in table 5.1 is challenging for researchers on their own also applies to a certain extent to the WOAP. Setting up a basic study is relatively easy without extensive IT skills. However, more elaborate customization cannot currently be achieved without altering the codebase. When using a web-based platform with observers scattered in locations, it can be challenging to account for and control variance components in the form of varying ambient lighting conditions and diagnostic displays. There is, however, evidence suggesting that there is a similar diagnostic performance using radiological workstation displays vs. off-the-shelf displays[86,87] .

The I and III papers reported measurement agreement in continuous variables compared to different measurement techniques and modalities, respectively. In contrast, IV sought to quantify measurement agreement and sources of variation without comparison to other techniques/modalities. All three studies performed repeated measurements on the same set of images for each session without access to prior measurements. Such a reading paradigm is not by any means unique to the three application studies, and it is, in fact, a commonly used approach, providing a surrogate for the threshold for detection of true biological change. In this regard, it is essential to acknowledge that quantification of measurement variability, such as those reported in IV, may prove to be conservative estimates of the variability expected for follow-up studies in clinical practice. Evaluating datasets where measured lesions have demonstrated interval growth and where access to previous images is given when placing calipers may mitigate measurement variability[88].

Future perspectives

Additional observer agreement studies using the developed WOAP are currently being planned in our research group. Two of these will investigate the feasibility of using a contrast-enhanced low-dose CT and non-contrast MRI protocol, respectively, for radiation dose-saving in active surveillance of small renal masses. A future goal of the WOAP is to provide a more well-structured- and documented codebase that will allow researchers outside of our research group to use the developed platform. The code for a prior version of the Mulrecon DICOM viewer is already released for download[89,90], and we intend to release the code for the back-end as well. However, creating a complete software package with proper documentation where a setup of an observer performance study with unique features can be completed without altering the code base is a huge endeavor. Developing the necessary content management system is beyond our immediate and near future resources. However, a lightweight and well-documented basic backend code template should go a long way towards facilitating observer agreement studies in other research groups. Interestingly, recently, an elective Data Science Pathway for 4th-year radiology residents has been described and piloted[91]. As part of this Data Science Pathway, residents are exposed to aspects of AI-machine learning application development, including achieving proficiency in basic coding. Given the availability of proper code documentation, it is to be expected that a number of current radiologists will possess the necessary basic coding skills to customize and utilize a platform such as the WOAP for conducting observer performance studies.

Beyond accessing observer performance as part of research studies, the WOAP may play a role in radiology departmental benchmarking[92] - especially when coupled with an interactive statistical module similar to one employed in III and IV for easy comparison of results and clarification of errors and inconsistency in observer assessments. Furthermore, there is a need for research studies that investigate the effect of various educational interventions on observer performance which the WOAP can help facilitate[93].

Conclusion

The overall aim of this thesis was to develop and apply a web-based platform for facilitating observer performance studies in imaging research. In fulfillment of this overall aim, six specific aims were stipulated. Three clinical application studies and one method study were completed. In summary, the platform developed proved to be helpful in conducting studies encompassing a scale and scope of observers as recommended in the literature. Through an iterative process, the WOAP was gradually refined from completing an application study involving the evaluation of caliper measurements with the display of static 2D images (I, study 1) to the incorporation of a DICOM viewer capable of handling volumetric datasets with extensive automated data collection (studies 2-3, III-IV) and integration of interactive statistical analysis (aim I).

The first application study found that the ITI method in assessing maximal aortic diameter using ultrasound had superior reproducibility compared to OTO and LTL (aim II).

The method study and subsequent extensions showed that it was possible to develop a web-based DICOM viewer for interactive visualization of volumetric DICOM datasets with functionality and real-time rendering speed comparable to desktop-based PACS systems (aim III).

The second application study found that non-contrast CT scans at ultra-low-dose levels are interchangeable with gold-standard CTA to assess abdominal aortic diameter using double oblique multiplayer reconstruction. Measurements can be completed timely and compatible with clinical practice (aim IV).

The third application study demonstrated that CT-based two-point pancreatic measurements are subject to substantial intra- and interobserver variability even among specialists. The findings question the implementation of two-point measurements as the basis for imaging scoring systems in chronic pancreatitis (aim V).

In addition, data from papers I, III-IV demonstrated that the usability of the WOAP and Mulrecon DICOM viewer was on par with products with well-recognized high usability ratings. Furthermore, the results of the performance evaluation of the WOAP with accompanying DICOM viewer in terms of download and rendering speed were

satisfactory for the completion of observer agreement studies, and participating observers completed observer tasks at a high completion rate (aim VI).

References

1. Booij, R., Budde, R. P. J., Dijkshoorn, M. L. & van Straten, M. Technological developments of X-ray computed tomography over half a century: User's influence on protocol optimization. *European Journal of Radiology* vol. 131 109261 (2020).
2. Strimbu, K. & Tavel, J. A. What are biomarkers? *Curr. Opin. HIV AIDS* **5**, 463–466 (2010).
3. Abramson, R. G. *et al.* Methods and challenges in quantitative imaging biomarker development. *Acad. Radiol.* **22**, 25–32 (2015).
4. Rogers, W. *et al.* Radiomics: from qualitative to quantitative imaging. *Br. J. Radiol.* **93**, 20190948 (2020).
5. Lucignani, G. & Neri, E. Integration of imaging biomarkers into systems biomedicine: a renaissance for medical imaging. *Clinical and Translational Imaging* vol. 7 149–153 (2019).
6. Brady, A. P. *et al.* Radiology in the Era of Value-based Healthcare: A Multi-Society Expert Statement from the ACR, CAR, ESR, IS3R, RANZCR, and RSNA. *Radiology* **298**, 486–491 (2021).
7. Fryback, D. G. & Thornbury, J. R. The efficacy of diagnostic imaging. *Med. Decis. Making* **11**, 88–94 (1991).
8. Crewson, P. E. Reader agreement studies. *AJR Am. J. Roentgenol.* **184**, 1391–1397 (2005).
9. Cohen, J. F. *et al.* STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open* **6**, e012799 (2016).
10. Beam, C. A., Baker, M. E., Paine, S. S., Sostman, H. D. & Sullivan, D. C. Answering unanswered questions: proposal for a shared resource in clinical diagnostic radiology

research. *Radiology* **183**, 619–620 (1992).

11. Obuchowski, N. A. & Zepp, R. C. Simple steps for improving multiple-reader studies in radiology. *AJR Am. J. Roentgenol.* **166**, 517–521 (1996).
12. Obuchowski, N. A. How many observers are needed in clinical studies of medical imaging? *AJR Am. J. Roentgenol.* **182**, 867–869 (2004).
13. Zhou, X.-H., Obuchowski, N. A. & McClish, D. K. *Statistical Methods in Diagnostic Medicine*. (John Wiley & Sons, 2014).
14. Leo, G. D. & Di Leo, G. Measurements in radiology: the need for high reproducibility. *Pediatric Radiology* vol. 45 32–34 (2015).
15. Han, D. *et al.* Influence of lung nodule margin on volume- and diameter-based reader variability in CT lung cancer screening. *Br. J. Radiol.* **91**, 20170405 (2018).
16. Erasmus, J. J. *et al.* Interobserver and intraobserver variability in measurement of non-small-cell carcinoma lung lesions: implications for assessment of tumor response. *J. Clin. Oncol.* **21**, 2574–2582 (2003).
17. Orton, L. P. *et al.* Variability in computed tomography diameter measurements of solid renal masses. *Abdom. Imaging* **39**, 533–542 (2014).
18. Shiraishi, J., Pesce, L. L., Metz, C. E. & Doi, K. Experimental Design and Data Analysis in Receiver Operating Characteristic Studies: Lessons Learned from Reports inRadiologyfrom 1997 to 2006. *Radiology* vol. 253 822–830 (2009).
19. Farzin, B. *et al.* Agreement studies in radiology research. *Diagn. Interv. Imaging* **98**, 227–233 (2017).
20. Bankier, A. A., Levine, D., Halpern, E. F. & Kressel, H. Y. Consensus Interpretation in Imaging Research: Is There a Better Way? *Radiology* vol. 257 14–17 (2010).

21. Ochodo, E. A. *et al.* Overinterpretation and misreporting of diagnostic accuracy studies: evidence of 'spin'. *Radiology* **267**, 581–588 (2013).

22. Jones, M., Dobson, A. & O'Brian, S. A graphical method for assessing agreement with the mean between multiple observers using continuous measures. *Int. J. Epidemiol.* **40**, 1308–1313 (2011).

23. McDonald, R. J. *et al.* The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. *Acad. Radiol.* **22**, 1191–1198 (2015).

24. Andriole, K. P. *et al.* Optimizing analysis, visualization, and navigation of large image data sets: one 5000-section CT scan can ruin your whole day. *Radiology* **259**, 346–362 (2011).

25. den Boer, L. *et al.* Volumetric image interpretation in radiology: scroll behavior and cognitive processes. *Adv. Health Sci. Educ. Theory Pract.* **23**, 783–802 (2018).

26. Dalrymple, N. C., Prasad, S. R., Freckleton, M. W. & Chintapalli, K. N. Introduction to the Language of Three-dimensional Imaging with Multidetector CT. *RadioGraphics* vol. 25 1409–1428 (2005).

27. Brambilla, M., Vassileva, J., Kuchcinska, A. & Rehani, M. M. Multinational data on cumulative radiation exposure of patients from recurrent radiological procedures: call for action. *Eur. Radiol.* **30**, 2493–2501 (2020).

28. Loose, R. W. *et al.* Radiation dose management systems—requirements and recommendations for users from the ESR EuroSafe Imaging initiative. *European Radiology* vol. 31 2106–2114 (2021).

29. Boone, J. M., Hendee, W. R., McNitt-Gray, M. F. & Seltzer, S. E. Radiation Exposure from CT Scans: How to Close Our Knowledge Gaps, Monitor and Safeguard Exposure—Proceedings and Recommendations of the Radiation Dose Summit, Sponsored by

NIBIB, February 24–25, 2011. *Radiology* vol. 265 544–554 (2012).

30. Kubo, T. Vendor free basics of radiation dose reduction techniques for CT. *European Journal of Radiology* vol. 110 14–21 (2019).
31. Paulo, G. *et al.* Diagnostic Reference Levels based on clinical indications in computed tomography: a literature review. *Insights Imaging* **11**, 96 (2020).
32. Jensen, C. T. *et al.* Detection of Colorectal Hepatic Metastases Is Superior at Standard Radiation Dose CT versus Reduced Dose CT. *Radiology* **290**, 400–409 (2019).
33. Keller, G., Afat, S., Ahrend, M.-D. & Springer, F. Diagnostic accuracy of ultra-low-dose CT for torsion measurement of the lower limb. *Eur. Radiol.* **31**, 3574–3581 (2021).
34. Fletcher, J. G. *et al.* Estimation of Observer Performance for Reduced Radiation Dose Levels in CT: Eliminating Reduced Dose Levels That Are Too Low Is the First Step. *Acad. Radiol.* **24**, 876–890 (2017).
35. Olsen, Ø. E. Why measure tumours? *Pediatr. Radiol.* **45**, 35–41 (2015).
36. Börjesson, S. *et al.* A software tool for increased efficiency in observer performance studies in radiology. *Radiat. Prot. Dosimetry* **114**, 45–52 (2005).
37. Looney, P. T., Young, K. C. & Halling-Brown, M. D. MEDXVIEWER: PROVIDING A WEB-ENABLED WORKSTATION ENVIRONMENT FOR COLLABORATIVE AND REMOTE MEDICAL IMAGING VIEWING, PERCEPTION STUDIES AND READER TRAINING. *Radiat. Prot. Dosimetry* **169**, 32–37 (2016).
38. Anwyl-Irvine, A., Dalmaijer, E. S., Hodges, N. & Evershed, J. K. Realistic precision and accuracy of online experiment platforms, web browsers, and devices. *Behav. Res. Methods* **53**, 1407–1425 (2021).
39. Rubin, G. D. Data explosion: the challenge of multidetector-row CT. *Eur. J. Radiol.* **36**,

74–80 (2000).

40. Mandal, T. & Jana, B. A Study on Risk Assessment in Information Security. *SSRN Electronic Journal* doi:10.2139/ssrn.3261593.

41. Congote, J. *et al.* Interactive visualization of volumetric data with WebGL in real-time. *Proceedings of the 16th International Conference on 3D Web Technology - Web3D '11* (2011) doi:10.1145/2010425.2010449.

42. Wanhainen, A. *et al.* Editor's Choice - European Society for Vascular Surgery (ESVS) 2019 Clinical Practice Guidelines on the Management of Abdominal Aorto-iliac Artery Aneurysms. *Eur. J. Vasc. Endovasc. Surg.* **57**, 8–93 (2019).

43. Zucker, E. J. & Prabhakar, A. M. Abdominal aortic aneurysm screening: concepts and controversies. *Cardiovasc Diagn Ther* **8**, S108–S117 (2018).

44. Frøkjær, J. B. *et al.* Guidelines for the Diagnostic Cross Sectional Imaging and Severity Scoring of Chronic Pancreatitis. *Pancreatology* **18**, 764–773 (2018).

45. Usage Statistics and Market Share of PHP for Websites, September 2021. https://w3techs.com/technologies/details/pl-php.

46. Williams, H. E. & Lane, D. *Web Database Applications with PHP and MySQL: Building Effective Database-Driven Web Sites*. ('O'Reilly Media, Inc.', 2004).

47. Abriata, L. A., Rodrigues, J. P. G. L. M., Salathé, M. & Patiny, L. Augmenting Research, Education, and Outreach with Client-Side Web Programming. *Trends Biotechnol.* **36**, 473–476 (2018).

48. Borgbjerg, J. *et al.* Superior Reproducibility of the Leading to Leading Edge and Inner to Inner Edge Methods in the Ultrasound Assessment of Maximum Abdominal Aortic Diameter. *Eur. J. Vasc. Endovasc. Surg.* **55**, 206–213 (2018).

49. Borgbjerg, J. MULRECON: A Web-based Imaging Viewer for Visualization of Volumetric Images. *Curr. Probl. Diagn. Radiol.* **48**, 531–534 (2019).

50. Graham, R. N. J., Perriss, R. W. & Scarsbrook, A. F. DICOM demystified: a review of digital file formats and their use in radiological practice. *Clin. Radiol.* **60**, 1133–1140 (2005).

51. Preim, B. & Botha, C. P. *Visual Computing for Medicine: Theory, Algorithms, and Applications*. (Newnes, 2013).

52. Guidelines and standards for implementation of new PACS/RIS solutions in the UK. https://www.rcr.ac.uk/publication/guidelines-and-standards-implementation-new-pacsris-solutions-uk.

53. Borgbjerg, J. Web-based imaging viewer for real-color volumetric reconstruction of human visible project and DICOM datasets. *Clin. Anat.* **34**, 470–477 (2021).

54. Ghayour, F. & Cantor, D. *Real-Time 3D Graphics with WebGL 2: Build interactive 3D applications with JavaScript and WebGL 2 (OpenGL ES 3.0), 2nd Edition*. (Packt Publishing Ltd, 2018).

55. Choudhri, A. F., Chatterjee, A. R., Javan, R., Radvany, M. G. & Shih, G. Security Issues for Mobile Medical Imaging: A Primer. *Radiographics* **35**, 1814–1824 (2015).

56. Guerrilla HCI: Article by Jakob Nielsen. https://www.nngroup.com/articles/guerrilla-hci/.

57. Borgbjerg, J. Novel web-based tool for conducting observer performance studies in imaging research. (2016).

58. Bangor, A., Kortum, P. T. & Miller, J. T. An Empirical Evaluation of the System Usability Scale. *International Journal of Human-Computer Interaction* vol. 24 574–594 (2008).

59. Lewis, J. R. The System Usability Scale: Past, Present, and Future. *International Journal of Human–Computer Interaction* vol. 34 577–590 (2018).

60. Kortum, P. T. & Bangor, A. Usability Ratings for Everyday Products Measured With the System Usability Scale. *International Journal of Human-Computer Interaction* vol. 29 67–76 (2013).

61. User Experience Magazine. Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale. https://uxpajournal.org/determining-what-individual-sus-scores-mean-adding-an-adjective-rating-scale/.

62. Sauro, J. *A Practical Guide to the System Usability Scale: Background, Benchmarks & Best Practices*. (CreateSpace, 2011).

63. Lisitskaya, M. V. *et al.* Systematic approach for assessment of imaging features in chronic pancreatitis: a feasibility and validation study from the Scandinavian Baltic Pancreatic Club (SBPC) database. *Abdom Radiol (NY)* **45**, 1468–1480 (2020).

64. Bland, J. M., Martin Bland, J. & Altman, D. STATISTICAL METHODS FOR ASSESSING AGREEMENT BETWEEN TWO METHODS OF CLINICAL MEASUREMENT. *The Lancet* vol. 327 307–310 (1986).

65. Anvari, A., Halpern, E. F. & Samir, A. E. Essentials of Statistical Methods for Assessing Reliability and Agreement in Quantitative Imaging. *Acad. Radiol.* **25**, 391–396 (2018).

66. Christensen, H. S., Borgbjerg, J., Børty, L. & Bøgsted, M. On Jones et al.'s method for extending Bland-Altman plots to limits of agreement with the mean for multiple observers. doi:10.21203/rs.3.rs-42805/v2.

67. Parker, R. A., Scott, C., Inácio, V. & Stevens, N. T. Using multiple agreement methods for continuous repeated measures data: a tutorial for practitioners. *BMC Medical Research Methodology* vol. 20 (2020).

68. Christensen, H. S., Borgbjerg, J., Børty, L. & Bøgsted, M. On Jones et al.'s Method for Assessing Limits of Agreement with the Mean for Multiple Observers. doi:10.21203/rs.3.rs-42805/v1.

69. Kakinuma, R. *et al.* Measurement of Focal Ground-glass Opacity Diameters on CT Images. *Academic Radiology* vol. 19 389–394 (2012).

70. Mohammadinejad, P. *et al.* CT Noise-Reduction Methods for Lower-Dose Scanning: Strengths and Weaknesses of Iterative Reconstruction Algorithms and New Techniques. *Radiographics* **41**, 1493–1508 (2021).

71. McErlean, A. *et al.* Intra- and Interobserver Variability in CT Measurements in Oncology. *Radiology* vol. 269 451–459 (2013).

72. Yang, H., Guo, X., Schwartz, L. H. & Zhao, B. A Web-Based Response-Assessment System for Development and Validation of Imaging Biomarkers in Oncology. *Tomography* **5**, 220–225 (2019).

73. Rubin, D. L., Ugur Akdogan, M., Altindag, C. & Alkim, E. ePAD: An Image Annotation and Analysis Platform for Quantitative Imaging. *Tomography* **5**, 170–183 (2019).

74. Hostetter, J., Khanna, N. & Mandell, J. C. Integration of a Zero-footprint Cloud-based Picture Archiving and Communication System with Customizable Forms for Radiology Research and Education. *Acad. Radiol.* **25**, 811–818 (2018).

75. Ziegler, E. *et al.* Open Health Imaging Foundation Viewer: An Extensible Open-Source Framework for Building Web-Based Imaging Applications to Support Cancer Research. *JCO Clin Cancer Inform* **4**, 336–345 (2020).

76. Genske, U. & Jahnke, P. Human Observer Net: A Platform Tool for Human Observer Studies of Image Data. *Radiology* 211832 (2022).

77. Thompson, J. D. Toward Consistent Design and Reporting of Observer Studies in Imaging. *Radiology* 220150 (2022).

78. Canellas, R. *et al.* Abbreviated MRI Protocols for the Abdomen. *Radiographics* **39**, 744–758 (2019).

79. Smith-Bindman, R. *et al.* An Image Quality–informed Framework for CT Characterization. *Radiology* (2021) doi:10.1148/radiol.2021210591.

80. Borgbjerg, J. *et al.* CT-guided cryoablation of renal cancer: radiation burden and the associated risk of secondary cancer from procedural- and follow-up imaging. *Abdom Radiol (NY)* **45**, 3581–3588 (2020).

81. Borgbjerg, J., Nilsen, F. S. & Nielsen, T. K. Unenhanced MRI for Surveillance of Small Solid Renal Masses: Additional Evidence Is Needed. *AJR. American journal of roentgenology* vol. 217 1017 (2021).

82. Mawi, H., Narine, R. & Schieda, N. Adequacy of Unenhanced MRI for Surveillance of Small (Clinical T1a) Solid Renal Masses. *AJR Am. J. Roentgenol.* **216**, 960–966 (2021).

83. Schmid, A. M. *et al.* Radiologists and Clinical Trials: Part 1 The Truth About Reader Disagreements. *Ther Innov Regul Sci* **55**, 1111–1121 (2021).

84. Allen, B., Agarwal, S., Coombs, L., Wald, C. & Dreyer, K. 2020 ACR Data Science Institute Artificial Intelligence Survey. *J. Am. Coll. Radiol.* **18**, 1153–1159 (2021).

85. Yang, L. *et al.* Stakeholders' perspectives on the future of artificial intelligence in radiology: a scoping review. *Eur. Radiol.* (2021) doi:10.1007/s00330-021-08214-z.

86. Doyle, A. J., Le Fevre, J. & Anderson, G. D. Personal computer versus workstation display: observer performance in detection of wrist fractures on digital radiographs. *Radiology* **237**, 872–877 (2005).

87. Bhatia, A. *et al.* Intra and Inter-Observer Reliability of Mobile Tablet PACS Viewer System vs. Standard PACS Viewing Station-Diagnosis of Acute Central Nervous System Events. *Open Journal of Radiology* vol. 03 91–98 (2013).

88. Bankier, A. A. *et al.* Recommendations for Measuring Pulmonary Nodules at CT: A Statement from the Fleischner Society. *Radiology* vol. 285 584–600 (2017).

89. Borgbjerg, J. Mulrecon. https://www.castlemountain.dk/atlas/index.php?page=mulrecon.

90. Borgbjerg, J. Mulrecon Color. https://www.castlemountain.dk/atlas/index.php?page=mulrecon&mulreconPage=color.

91. Wiggins, W. F. *et al.* Preparing Radiologists to Lead in the Era of Artificial Intelligence: Designing and Implementing a Focused Data Science Pathway for Senior Radiology Residents. *Radiol Artif Intell* **2**, e200057 (2020).

92. Mahgerefteh, S., Kruskal, J. B., Yam, C. S., Blachar, A. & Sosna, J. Peer Review in Diagnostic Radiology: Current State and a Vision for the Future. *RadioGraphics* vol. 29 1221–1231 (2009).

93. Woo, M., Lowe, S. C., Devane, A. M. & Gimbel, R. W. Intervention to Reduce Interobserver Variability in Computed Tomographic Measurement of Cancer Lesions Among Experienced Radiologists. *Curr. Probl. Diagn. Radiol.* **50**, 321–327 (2021).

94. Netravali, A. N. *Digital Pictures: Representation, Compression, and Standards*. (Springer, 2013).

95. Williams, L. H. & Drew, T. What do we know about volumetric medical image interpretation?: a review of the basic science and medical image perception literatures. *Cogn Res Princ Implic* **4**, 21 (2019).

96. Diaz, I., Schmidt, S., Verdun, F. R. & Bochud, F. O. Eye-tracking of nodule detection in lung

CT volumetric data. *Medical Physics* vol. 42 2925–2932 (2015).

97. Guidelines and standards for implementation of new PACS/RIS solutions in the UK. https://www.rcr.ac.uk/publication/guidelines-and-standards-implementation-new-pacsris-solutions-uk.

Appendix A

An essential parameter in DICOM viewer performance is reviewing volumetric imaging datasets with a sufficiently high cine scroll speed. The preferred cine scroll speed varies considerably among radiologists for different imaging modalities and diagnostic tasks. However, basic research in visual psychophysics has shown that the optimal temporal frequency for contrast and motion sensitivity is between 4 and 16 Hertz[94], which is correlated to the frame rates at which radiologists scroll through image stacks[95]. In terms of scroll speed used in clinical practice, one study evaluating CT-based lung cancer detection in which large volumes of data need to be covered in a short time period of time found that most readers employed a speed around 25-30 frames per second[96]. An experiment was conducted to evaluate the developed GPU-accelerated Mulrecon DICOM viewer employed in application studies 2 and 3 in terms of speed. A CT dataset from application study 2 was used for this experiment. The dataset had dimensions of 512x512x200 voxels. The speed tests were conducted using a laptop system from 2016 with the following hardware specifications: Intel core i7-8750H 6 Core CPU at 2.20 GHz with 8GB of RAM and a GeForce GTX 1060 running Windows 10 64 bits. For this experiment, among available web browsers with full implementation of the WebGL 2.0 standard utilized in the Mulrecon DICOM viewer, we selected the following: Firefox, Chrome, Opera, and Microsoft Edge.

We used an empirical speed measuring method. First, the axial stack of the Mulrecon DICOM viewer was set to a slice thickness of 5 mm. We then forced the DICOM viewer to continuously scroll the axial stack and counted how many times the stack was updated in a 10-seconds interval. This procedure was repeated five times, and the average rendering scroll speed was determined. Chrome, Firefox, Edge, and Opera demonstrated a scroll speed of 26, 22, 32, and 31 frames per second, respectively, yielding an average of 28 frames per second among the four browsers. Hence, the achieved cine scroll speed is comparable to the scroll speed referenced above.

Appendix B

To the editor,

In our original article *Superior Reproducibility of the Leading to Leading Edge and Inner to Inner Edge Methods in the Ultrasound Assessment of Maximum Abdominal Aortic Diameter,* published 2018 in European Journal of Vascular and Endovascular surgery, we reported on observer reproducibility of caliper placement in ultrasonographic determination of maximum abdominal aortic diameter with the three principal methods: leading to leading edge (LTL), inner to inner edge (ITI), and outer to outer edge (OTO)[48]. We concluded that the LTL and ITI have superior reproducibility compared with the OTO method. However, a corrigendum is needed.
First, two numbers in the "Measurements" sub-section of the "Materials and methods" section are incorrect. The sentence "...a total of 1350 measurements and 450 caliper placements for each of the three methods…" should have read "...a total of 1800 measurements and 600 caliper placements for each of the three methods".
In addition, in 2020, Christensen et al. have provided an extension of this model with interobserver variance incorporated, and thus reproducibility[66]. In this update, *limits of agreement* can be defined as *how much a given observer's measurement may plausibly deviate from the mean of all observers' measurements on the specific subject*. Hence, it has become clear/apparent that the limits of agreement with the mean (LOAM) statistical method used in our paper from 2018 to estimate observer agreement does not fully integrate variation due to different observers, and therefore the reported results do not properly reflect reproducibility. In our paper, single and repeated measurements were performed by 18 and a subset of 12 observers, respectively. Based on this new extension of the model[2], we in this corrigendum present LOAMs with accompanying variance components derived from these data in table 1. The updated LOAMs from single measurements are in alignment with the original conclusion of superior reproducibility of LTL and ITI compared to OTO. However, the updated results from repeated measurements demonstrate the greatest inter-observer variance component with respect to LTL. Nonetheless, the residual variance which approximates measurement repeatability is the smallest for LTL and ITI across single as well as repeated measurements. In all, these updated results somewhat question the original conclusion of superiority of LTL compared to OTO in terms of reproducibility. It does, however, solidify the original recommendation to adopt the ITI method as part of a standard procedure in ultrasound

assessment of abdominal aortic size.

Table 1. Updated LOAM statical analysis according to Christensen et al., 2020: 95% limits of agreement with the mean, and inter-subject (**σ_A**), inter-observer (**σ_B**) as well as residual (**σ_E**) variance component estimates in mm for OTO, ITI, and LTL.

	LOAM (95% CI)	σ_A (95% CI)	σ_B (95% CI)	σ_E (95% CI)
Single measurements (n=18)				
OTO	3.4 (3.0 to 4.2)	7.2 (5.7 to 8.6)	1.13 (0.74 to 1.5)	1.35 (1.3 to 1.4)
LTL	2.7 (2.4 to 3.5)	6.9 (5.5 to 8.3)	1.1 (0.7 to 1.4)	0.9 (0.9 to 1.0)
ITI	2.7 (2.4 to 3.6)	6.7 (5.4 to 8.0)	1.1 (0.7 to 1.4)	0.9 (0.9 to 1.0)
Repeated measurements (n=12)				
OTO	3.2 (2.8 to 4.3)	7.2 (5.7 to 8.6)	1.1 (0.7 to 1.6)	1.2 (1.2 to 1.3)
LTL	3.4 (2.8 to 5.1)	6.9 (5.5 to 8.3)	1.5 (0.8 to 2.1)	1.0 (1.0 to 1.1)
ITI	2.9 (2.4 to 4.3)	6.8 (5.4 to 8.1)	1.2 (0.7 to 1.8)	0.9 (0.9 to 0.9)

Appendix C

Guidelines and standards for implementation of new PACS/RIS solutions in the UK - CT display requirements

Regarding display of volumetric imaging the guidelines state "Increasing volumes of data are being produced from CT examinations such as in cardiology, CT colonography, oncology and trauma. The volumes of data consist of increasingly thin slice thicknesses which may require manipulation as in MPR or 3D reformatting. The viewing application must have the following functions"[97].

	Requirement	Implemented in the Mulrecon imaging viewer
1.	The PACS image display application must have an automatic and seamless loading of MPR to manipulate the thin CT slices rather than reliance on stand-alone modality workstations which are time consuming and inefficient. This facility will negate the need to store to PACS images in three orthogonal planes and therefore reduce the pressure on storage capacity.	X
2	Allow for synchronized scrolling in 3 planes for cross-sectional imaging (CT/MRI).	X
3	Allow automatic display of relevant prior	X
4	Allow synchronised scrolling with previous scan	X
5	Ability to create 3D images.	X
6	During MPR/3D viewing, radiologists should be able to save some key images (for example, a coronal image/sagittal image that shows the key lesion) as a	X

	separate series for reference to the report.	
7	Users must be able to define slab thickness and create images of different thickness in real time.	X
8	Ability to measure distance, circumference, angle and volume of lesions, this should be easy and intuitive.	X
9	Ability to measure Hounsfield density (for example, average density of a lung nodule, with maximum and minimum density). This task should be intuitive and easy for any radiologist.	X
10	Scrolling speed should be such that image to image transition is smooth – even with >1000 images. The users should be able to scroll through images smoothly. Cine display must be present.	X
11	CT displays at home-based applications for on-call reporting radiologists may be inadequate due to limited bandwidth. Scrolling speed over slow networks may be restricted but can be improved by the provision of local caching.	X

Thesis papers I-IV

I. **Borgbjerg J**, Bøgsted M, Lindholt JS, Behr-Rasmussen C, Hørlyck A, Frøkjær JB. Superior Reproducibility of the Leading to Leading Edge and Inner to Inner Edge Methods in the Ultrasound Assessment of Maximum Abdominal Aortic Diameter. Eur J Vasc Endovasc Surg. 2018 Feb;55(2):206-213. doi: 10.1016/j.ejvs.2017.11.019. Epub 2017 Dec 23. PMID: 29277483.

II. **Borgbjerg J**. MULRECON: A Web-based Imaging Viewer for Visualization of Volumetric Images. Curr Probl Diagn Radiol. 2019 Nov-Dec;48(6):531-534. doi: 10.1067/j.cpradiol.2018.09.001. Epub 2018 Sep 19. PMID: 30340913.

III. **Borgbjerg J**. Christensen HS, Al-Mashhadi R, Bøgsted M, Froekjaer J, Medrud L, Larsen N, Pedersen M, Thygesen J, Sivesgaard K, Lindholt J. Ultra-low-dose non-contrast CT is an adequate replacement for CT angiography in the assessment of maximal abdominal aortic diameter - [resubmitted (major revision) to Acta Radiologica Open]

IV. **Borgbjerg J**. Steinkohl E, Olesen S, Akisik F, Bethke A, Bieliuniene E, Christensen H, Engjom T, Haldorsen I, Kartalis N, Lisitskaya M, Naujokaite G, Novovic S, Ozola-Zālīte I, Phillips A, Swensson J, Drewes A, Frøkjær J. Observer Variability of Parenchymal- and Ductal Diameters in Chronic Pancreatitis: A Multi-institutional Study of CT Images - [resubmitted to Abdominal Radiology]

ISSN (online): 2246-1302
ISBN (online): 978-87-7573-843-4

AALBORG UNIVERSITY PRESS