Der Einfluss diagnostischer Empfehlungen auf die Entscheidungsfindung von Mediziner*innen

Präsentation am 21.09.2022 im Rahmen der Veranstaltung “Zukunft Teleradiologie“

Dr. Susanne Gaube
LMU München & Universitätsklinikum Regensburg

Funded by
VolkswagenStiftung

nature medicine

LETTERS

https://doi.org/10.1038/s41591-019-0447-x

End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography

Diego Ardila[1,5], Atilla P. Kiraly[1,5], Sujeeth Bharadwaj[1,5], Bokyung Choi[1,5], Joshua J. Reicher[2], Lily Peng[1], Daniel Tse[1*], Mozziyar Etemadi[3], Wenxing Ye[1], Greg Corrado[1], David P. Naidich[4] and Shravya Shetty[1]

With an estimated 160,000 deaths in 2018, lung cancer is the most common cause of cancer death in the United States[1]. Lung cancer screening using low-dose computed tomography has been shown to reduce mortality by 20–43% and is now included in US screening guidelines[1–6]. Existing challenges include inter-grader variability and high false-positive and false-negative rates[7–10]. We propose a deep learning algorithm that uses a patient's current and prior computed tomography volumes to predict the risk of lung cancer. Our model achieves a state-of-the-art performance (94.4% area under the curve) on 6,716 National Lung Cancer Screening Trial cases, and performs similarly on an independent clinical validation set of 1,139 cases. We conducted two reader studies. When prior computed tomography imaging was not available, our model outperformed all six radiologists with absolute reductions of 11% in false positives and 5% in false negatives. Where prior computed tomography imaging was available, the model performance was on-par with the same radiologists. This creates an opportunity to optimize the screening process via computer assistance and automation. While the vast majority of patients remain unscreened, we show the potential for deep learning models to increase the accuracy, consistency and adoption of lung cancer screening worldwide.

In 2013, the United States Preventive Services Task Force recommended low-dose computed tomography (LDCT) lung cancer screening in high-risk populations based on reported improved mortality in the National Lung Cancer Screening Trial (NLST)[2–5]. In 2014, the American College of Radiology published the Lung-RADS guidelines for LDCT lung cancer screening, to standardize image interpretation by radiologists and dictate management recommendations[1,6]. Evaluation is based on a variety of image findings, but primarily nodule size, density and growth[6]. At screening sites, Lung-RADS and other models such as PanCan are used to determine malignancy risk ratings that drive recommendations for clinical management[11,12]. Improving the sensitivity and specificity of lung cancer screening is imperative because of the high clinical and financial costs of missed diagnosis, late diagnosis and unnecessary biopsy procedures resulting from false negatives and false positives[5,13–17]. Despite improved consistency, persistent inter-grader variability and incomplete characterization of comprehensive imaging findings remain as limitations[7–10] of Lung-RADS. These limitations suggest opportunities for more sophisticated systems to improve performance and inter-reader consistency[18,19]. Deep learning approaches offer the exciting potential to automate more complex image analysis, detect subtle holistic imaging findings and unify methodologies for image evaluation[20].

A variety of software devices have been approved by the Food and Drug Administration (FDA) with the goal of addressing workflow efficiency and performance through augmented detection of lung nodules on lung computed tomography (CT)[21]. Clinical research has primarily focused on either nodule detection or diagnostic support for lesions manually selected by imaging experts[22–27]. Nodule detection systems were engineered with the goal of improving radiologist sensitivity in identifying nodules while minimizing costs to specificity, thereby falling into the category of computer-aided detection (CADe)[28]. This approach highlights small nodules, leaving malignancy risk evaluation and clinical decision making to the clinician. Diagnostic support for pre-identified lesions is included in computer-aided diagnosis (CADx) platforms, which are primarily aimed at improving specificity. CADx has gained greater interest and even first regulatory approvals in other areas of radiology, though not in lung cancer at the time of manuscript preparation[29].

To move beyond the limitations of prior CADe and CADx approaches, we aimed to build an end-to-end approach performing both localization and lung cancer risk categorization tasks using the input CT data alone. More specifically, we were interested in replicating a more complete part of a radiologist's workflow, including full assessment of LDCT volume, focus on regions of concern, comparison to prior imaging when available and calibration against biopsy-confirmed outcomes.

Another important high-level decision in our approach was to learn features using deep convolutional neural networks (CNN), rather than using hand-engineered features such as texture features or specific Hounsfield unit values. We chose to learn features because this approach has repeatedly been shown superior to hand-engineered features in many open computer vision competitions in the past five years[30,31], including the Kaggle 2017 Data Science Bowl which used NLST data[32].

There were three key components in our new approach (Fig. 1). First, we constructed a three-dimensional (3D) CNN model that performs end-to-end analysis of whole-CT volumes, using LDCT

doi:10.1038/nature21056

ification of skin cancer
:s

Susan M. Swetter[2,4], Helen M. Blau[5] & Sebastian Thrun[6]

images (for example, smartphone images) exhibit variability in factors such as zoom, angle and lighting, making classification substantially more challenging[23,24]. We overcome this challenge by using a data-driven approach—1.41 million pre-training and training images make classification robust to photographic variability. Many previous techniques require extensive preprocessing, lesion segmentation and extraction of domain-specific visual features before classification. By contrast, our system requires no hand-crafted features; it is trained end-to-end directly from image labels and raw pixels, with a single network for both photographic and dermoscopic images. The existing body of work uses small datasets of typically less than a thousand images of skin lesions[16,18,19], which, as a result, do not generalize well to new images. We demonstrate generalizable classification with a new dermatologist-labelled dataset of 129,450 clinical images, including 3,374 dermoscopy images.

Deep learning algorithms, powered by advances in computation and very large datasets[25], have recently been shown to exceed human performance in visual tasks such as playing Atari games[26], strategic board games like Go[27] and object recognition[6]. In this paper we outline the development of a CNN that matches the performance of dermatologists at three key diagnostic tasks: melanoma classification, melanoma classification using dermoscopy and carcinoma classification. We restrict the comparisons to image-based classification.

We utilize a GoogleNet Inception v3 CNN architecture[9] that was pre-trained on approximately 1.28 million images (1,000 object categories) from the 2014 ImageNet Large Scale Visual Recognition Challenge[6], and train it on our dataset using transfer learning[28]. Figure 1 shows the working system. The CNN is trained using 757 disease classes. Our dataset is composed of dermatologist-labelled images organized in a tree-structured taxonomy of 2,032 diseases, in which the individual diseases form the leaf nodes. The images come from 18 different clinician-curated, open-access online repositories, as well as from clinical data from Stanford University Medical Center. Figure 2a shows a subset of the full taxonomy, which has been organized clinically and visually by medical experts. We split our dataset into 127,463 training and validation images and 1,942 biopsy-labelled test images.

To take advantage of fine-grained information contained within the taxonomy structure, we develop an algorithm (Extended Data Table 1) to partition diseases into fine-grained training classes (for example, amelanotic melanoma and acrolentiginous melanoma). During inference, the CNN outputs a probability distribution over these fine classes. To recover the probabilities for coarser-level classes of interest (for example, melanoma) we sum the probabilities of their descendants (see Methods and Extended Data Fig. 1 for more details).

We validate the effectiveness of the algorithm in two ways, using nine-fold cross-validation. First, we validate the algorithm using a three-class disease partition—the first-level nodes of the taxonomy, which represent benign lesions, malignant lesions and non-neoplastic

nt of Dermatology, Stanford University, Stanford, California, USA. [3]Department of Pathology, ealth Care System, Palo Alto, California, USA. [5]Baxter Laboratory for Stem Cell Biology, Department

earning Algorithms
astases

); Bram van Ginneken, PhD;
the CAMELYON16 Consortium

athology images can

ithms at detecting
nodes of women with
tic setting.

on (CAMELYON16) to
lovember
2 centers in the
rified by
s to build algorithms.
9 whole-slide images
ding glass slides was
) from the Netherlands
-hour session,
ime constraint (WOTC).

e competition or

oci and the absence vs
rating characteristic
rated their diagnostic
mor, or definitely tumor.

AUC) for the algorithms
a lesion-level,
72.4% [95% CI,
slide image. For the
[95% CI, 0.983-0.999])
tic simulation (mean
a mean AUC that was
e of time constraints
.966 [95% CI,

on, some deep learning
pathologists
logy workflow;
terpreting whole-slide

Editorial page 2184

Related articles page 2211 and page 2250

Supplemental content

CME Quiz at jamanetwork.com/learning and CME Questions page 2252

Author Affiliations: Diagnostic Image Analysis Group, Department of Radiology and Nuclear Medicine, Radboud University Medical Center, Nijmegen, the Netherlands (Ehteshami Bejnordi, van Ginneken, Karssemeijer); Medical Image Analysis Group, Eindhoven University of Technology, Eindhoven, the Netherlands (Veta); Department of Pathology, University Medical Center Utrecht, Utrecht, the Netherlands (Johannes van Diest); Department of Pathology, Radboud University Medical Center, Nijmegen, the Netherlands (Litjens, van der Laak).

Group Information: The CAMELYON16 Consortium authors and collaborators are

'ERY

ep Learning Algorithm
y

Wu, BS; Arunachalam Narayanaswamy, PhD;
, PhD; Ramasamy Kim, OD, DNB;
D

an algorithm to
desired
nethods to

ection of diabetic

classification
development
: retinopathy,
ophthalmologists
resultant
sets, both
r consistency.

hm for detecting
c retinopathy,
erence standard
luated at 2
pecificity and

(mean age, 54.4
6]); the
2.6% women;
the algorithm
for EyePACS-1 and
with high
the specificity
5% CI, 81.1%-
erating point
7.5% and
y was 93.9%.

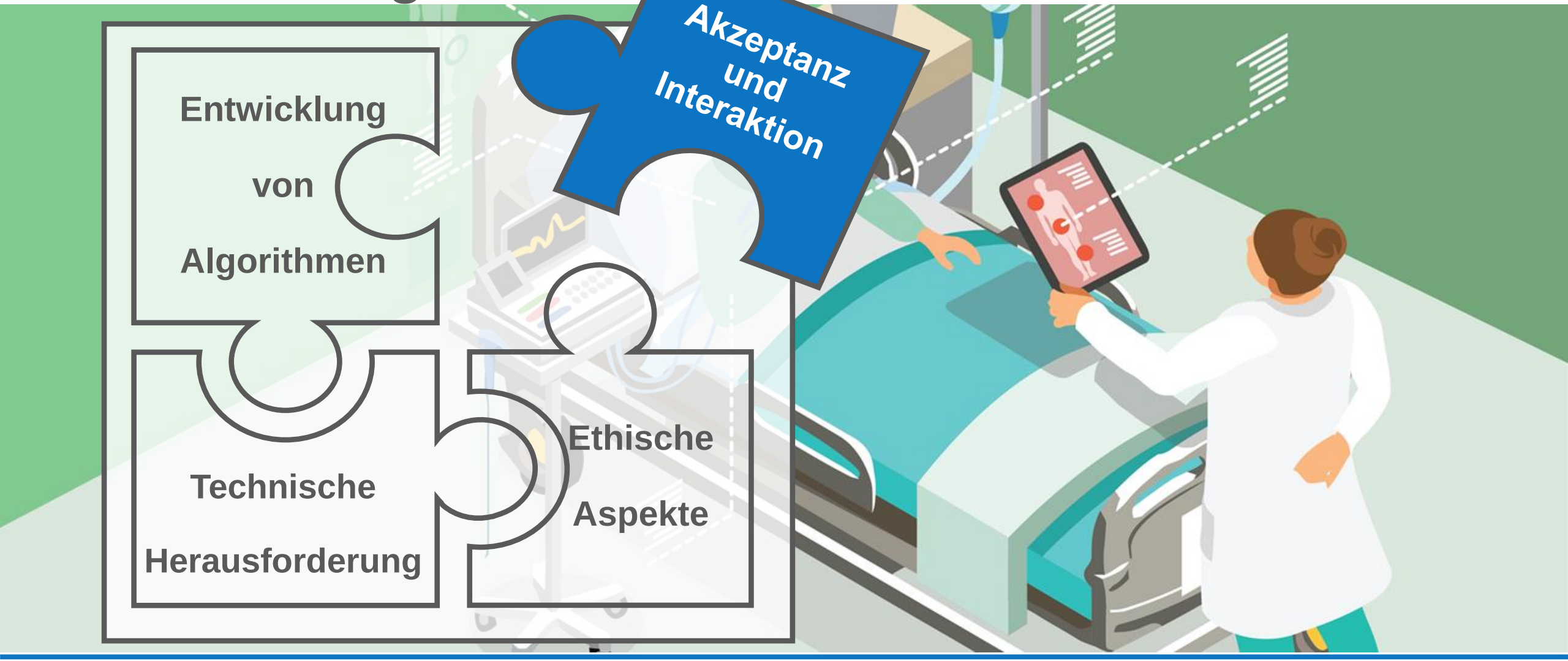
hs from adults
vity and
cessary to

Editorial pages 2366 and 2368

Supplemental content

Author Affiliations: Google Inc, Mountain View, California (Gulshan, Peng, Coram, Stumpe, Wu, Narayanaswamy, Venugopalan, Widner, Madams, Nelson, Webster); Department of Computer Science, University of Texas, Austin (Venugopalan); EyePACS LLC, San Jose, California (Cuadros); School of Optometry, Vision Science Graduate Group, University of California, Berkeley (Cuadros); Aravind Medical Research Foundation, Aravind Eye Care System, Madurai, India (Kim); Shri Bhagwan Mahavir Vitreoretinal Services, Sankara Nethralaya, Chennai, Tamil Nadu, India (Raman); Verily Life Sciences, Mountain View, California (Mega); Cardiovascular Division, Department of Medicine, Brigham and Women's Hospital and

Forschungsfelder



Forschungsfragen

Experiment 1

RQ1: Beeinflusst die …

1. **Quelle der Empfehlung** (KI vs. Mensch)
2. **Korrektheit der Empfehlung**
3. **Interaktion** zwischen Quelle und Korrektheit

die Performanz der Ärzte und ihre Bewertung der Empfehlungsqualität?

RQ2: Wirken sich Quelle und Korrektheit der Empfehlung auf **Aufgabenexperten** und **Nicht-Experten unterschiedlich** aus?

Experiment 2

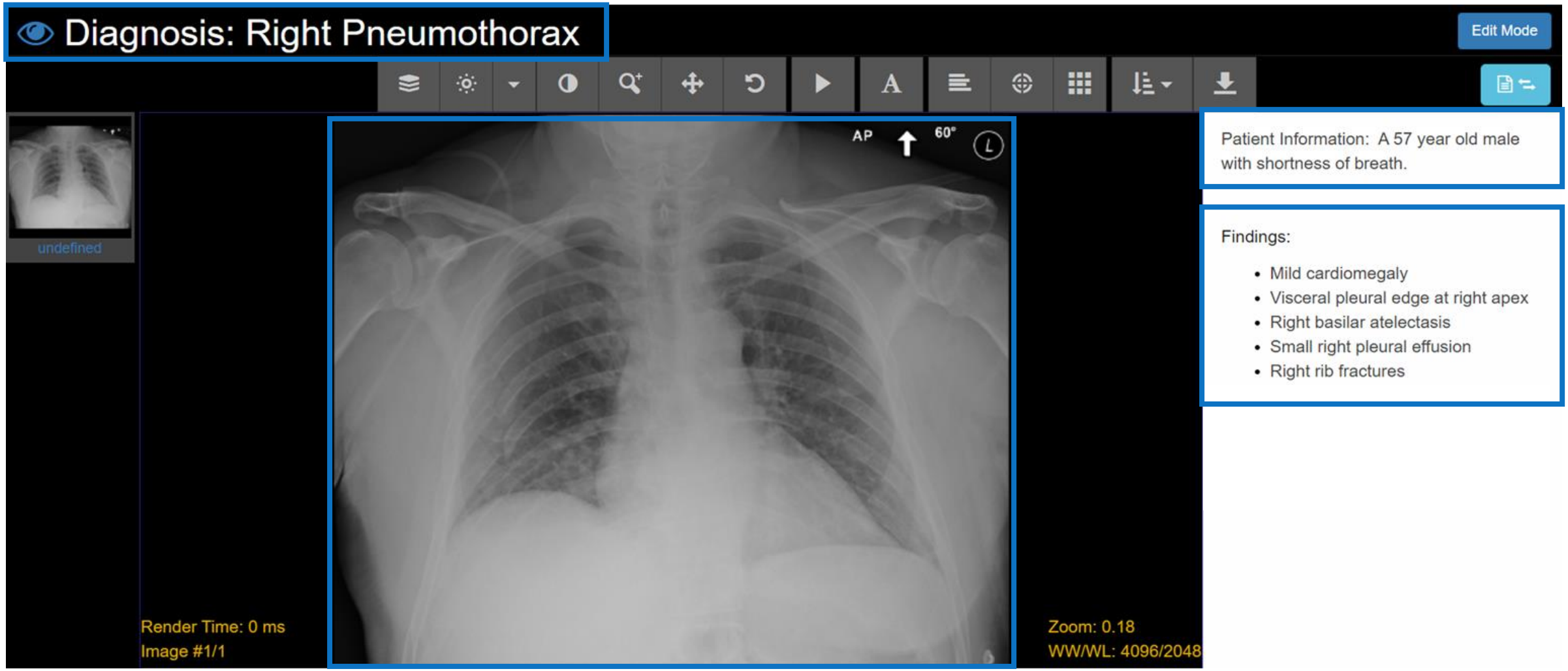
RQ1: Beeinflusst…

1. die **Quelle der Empfehlung** (KI vs. Mensch)
2. der **Grad der lokalen Erklärbarkeit**
3. die **Interaktion** zwischen Quelle und Erklärbarkeit

die Performanz der Ärzte, ihre Bewertung der Empfehlungsqualität und das Vertrauen in die eigene Entscheidung?

RQ2: Wirken sich Quelle und Grad der lokalen Erklärbarkeit der Empfehlung auf **Aufgabenexperten** und **Nichtexperten unterschiedlich** aus?

Methode Experiment 1



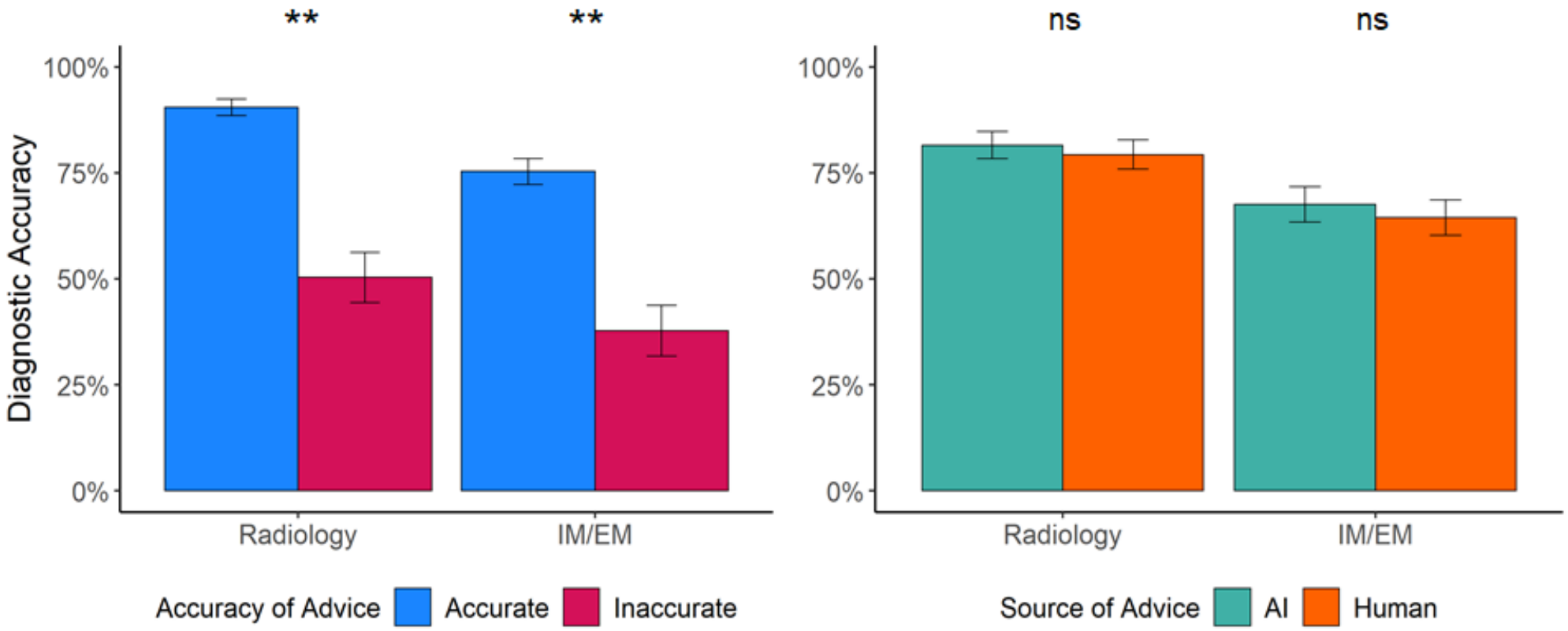
Methode Experiment 1

Übersicht
- Online-Experiment
- 138 Radiologen*innen
- 127 Internisten*innen/ Notfallmediziner*innen
- Design
  - Quelle der Empfehlung: Zwischensubjektfaktor
  - Korrektheit der Empfehlung: Innersubjektfaktor

Abhängige Variablen
- Performanz in %
- Bewertung der Empfehlungsqualität
  - Zustimmung
  - Nützlichkeit
  - Vertrauenswürdigkeit
  - Zukünftige Nutzung

Kontrollvariablen
- Berufliche Identifikation
- Berufliche Autonomie
- Einstellung zu KI
- Wissen zu KI
- Erfahrung in Jahren
- Geschlecht

Ergebnisse Experiment 1

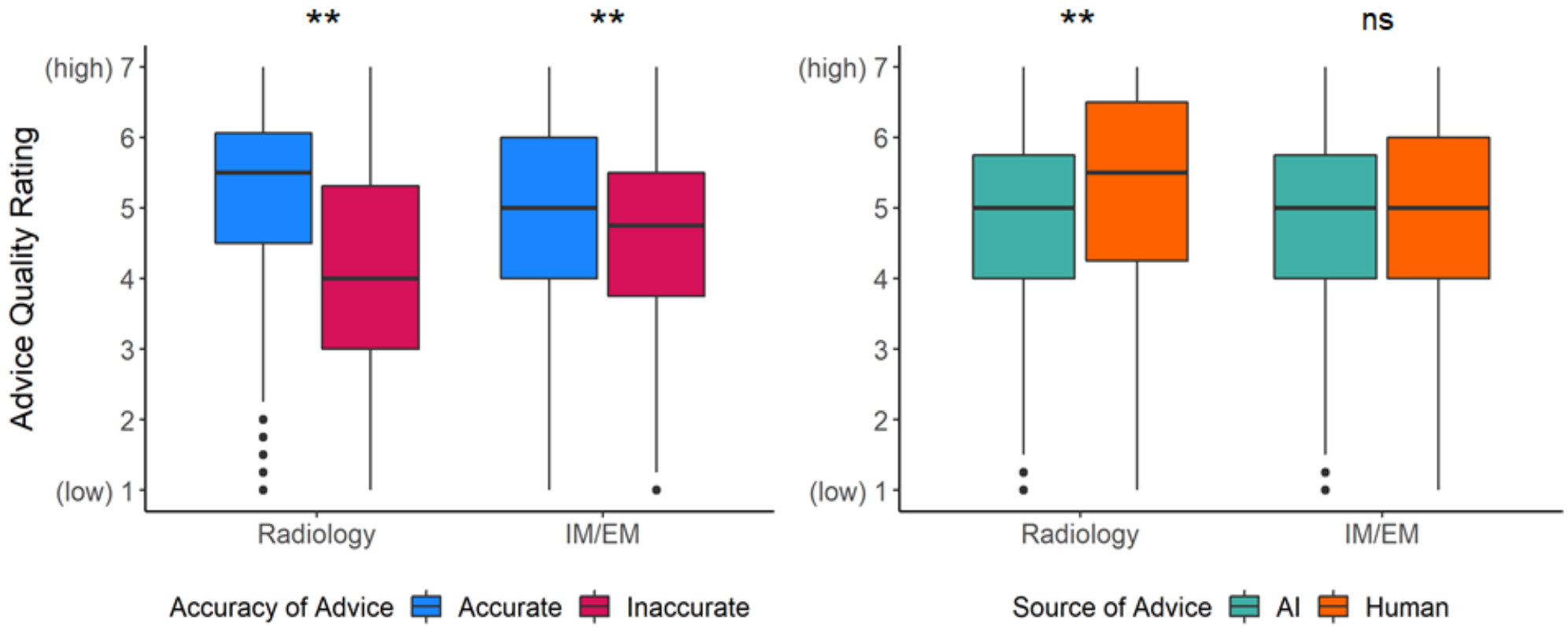


Logistic mixed multilevel regression models for diagnostic accuracy

	Task Experts (Radiology)				Non-Task Experts (IM/EM)			
Model 1	β	*SE*	*z*	*p*	β	*SE*	*z*	*p*
Intercept	-2.41	0.18	-13.68	< 0.001	-1.30	0.13	-9.71	< 0.001
Accuracy of the advice	2.39	0.24	9.79	< 0.001	1.91	0.23	8.26	< 0.001
Source of the advice	0.32	0.24	1.33	0.183	0.31	0.18	1.75	0.081
Accuracy x Source	-0.29	0.34	-0.85	0.394	-0.51	0.31	-1.62	0.105
Model 2 (incl. covariates)								
Intercept	-0.17	0.81	-0.21	0.835	-1.20	0.70	-1.71	0.087
Accuracy of the advice	2.43	0.25	9.86	< 0.001	1.91	0.23	8.26	< 0.001
Source of the advice	0.30	0.24	1.23	0.217	0.29	0.18	1.63	0.104
Accuracy x Source	-0.27	0.34	-0.78	0.436	-0.50	0.31	-1.61	0.108
Identification	-0.09	0.10	-0.95	0.342	0.01	0.08	0.08	0.937
Autonomy	-0.15	0.08	-2.03	0.042	-0.02	0.07	-0.28	0.780
AI Knowledge	-0.20	0.13	-1.55	0.120	-0.08	0.10	-0.76	0.445
AI Attitude	-0.07	0.08	-0.87	0.383	-0.02	0.08	-0.29	0.771
Experience	-0.04	0.01	-2.97	0.003	0.00	0.01	0.00	0.996
Gender	0.01	0.18	0.06	0.950	0.23	0.13	1.74	0.082

Note. β = estimated coefficient; *SE* = standard error; *z* = z-value; *p* = probability of committing a Type I error, IM = internal medicine, EM = emergency medicine.

Ergebnisse Experiment 1



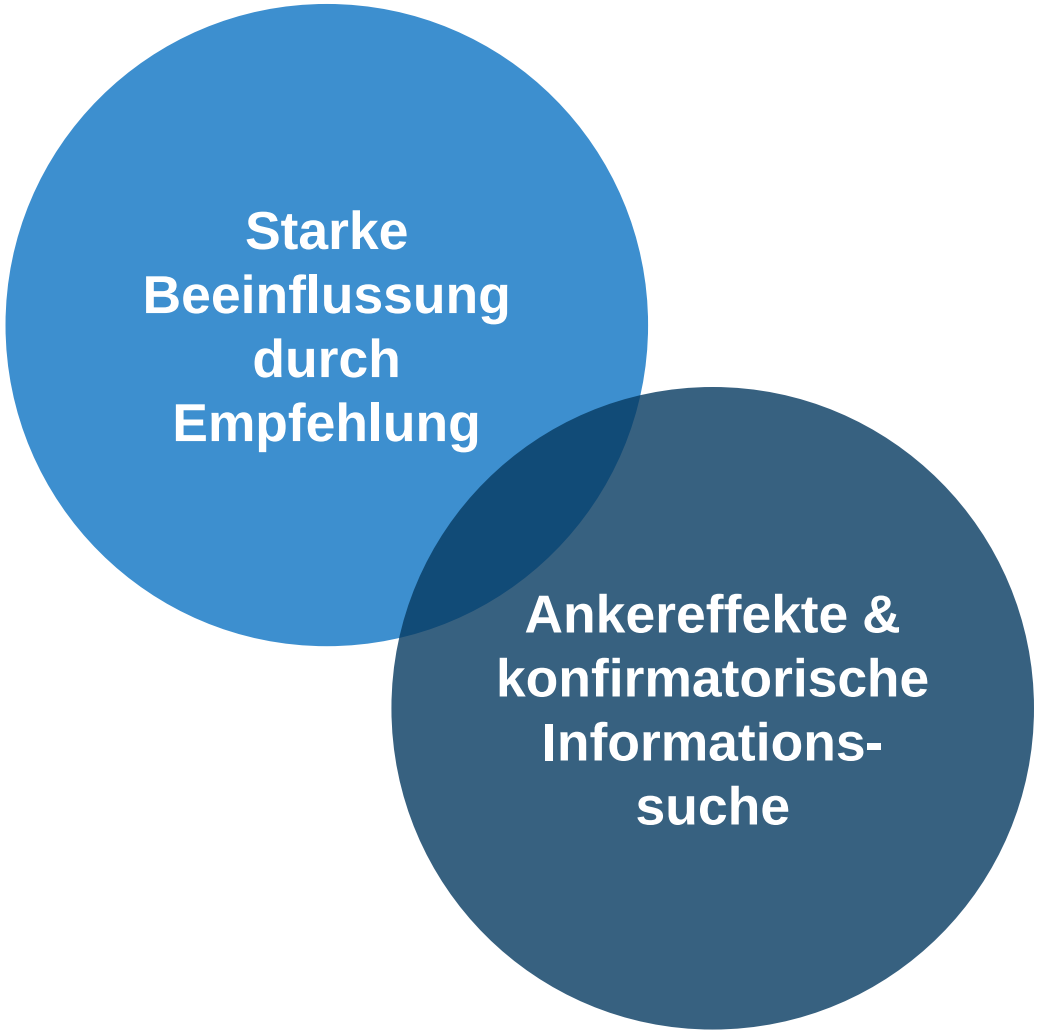
Linear mixed multilevel regression models for advice quality ratings

	Task Experts (Radiology)				Non-Task Experts (IM/EM)			
Model 1	γ	*SE*	*t*	*p*	γ	*SE*	*t*	*p*
Intercept	5.03	0.10	48.46	< 0.001	4.85	0.11	43.40	< 0.001
Accuracy of the advice	-1.00	0.12	-8.66	< 0.001	-0.40	0.12	-3.25	0.002
Source of the advice	0.53	0.15	3.5ᴄ	0.001	0.22	0.16	1.40	0.165
Accuracy x Source	-0.15	0.17	-0.90	0.368	-0.12	0.17	-0.73	0.469
Model 2 (incl. covariates)								
Intercept	1.74	0.56	3.08	0.002	5.06	0.67	7.55	< 0.001
Accuracy of the advice	-1.00	0.11	-8.98	< 0.001	-0.40	0.12	-3.25	0.001
Source of the advice	0.55	0.13	4.17	< 0.001	0.23	0.15	1.48	0.141
Accuracy x Source	-0.15	0.16	-0.94	0.351	-0.12	0.17	-0.73	0.469
Identification	0.27	0.07	4.00	< 0.001	0.03	0.07	0.46	0.646
Autonomy	0.12	0.06	2.05	0.043	0.04	0.07	0.61	0.545
AI Knowledge	0.21	0.09	2.33	0.021	-0.06	0.10	-0.61	0.542
AI Attitude	0.25	0.06	4.27	< 0.001	0.04	0.08	0.47	0.636
Experience	0.01	0.01	1.64	0.104	0.00	0.01	-0.19	0.849
Gender	-0.20	0.12	-1.65	0.102	-0.40	0.13	-3.11	0.002

Note. γ = regression coefficient; *SE* = standard error; *t* = t-value; *p* = probability of committing a Type I error, IM = internal medicine, EM = emergency medicine.

https://doi.org/10.1038/s41746-021-00385-9

Diskussion: Experiment 1



Methode Experiment 2

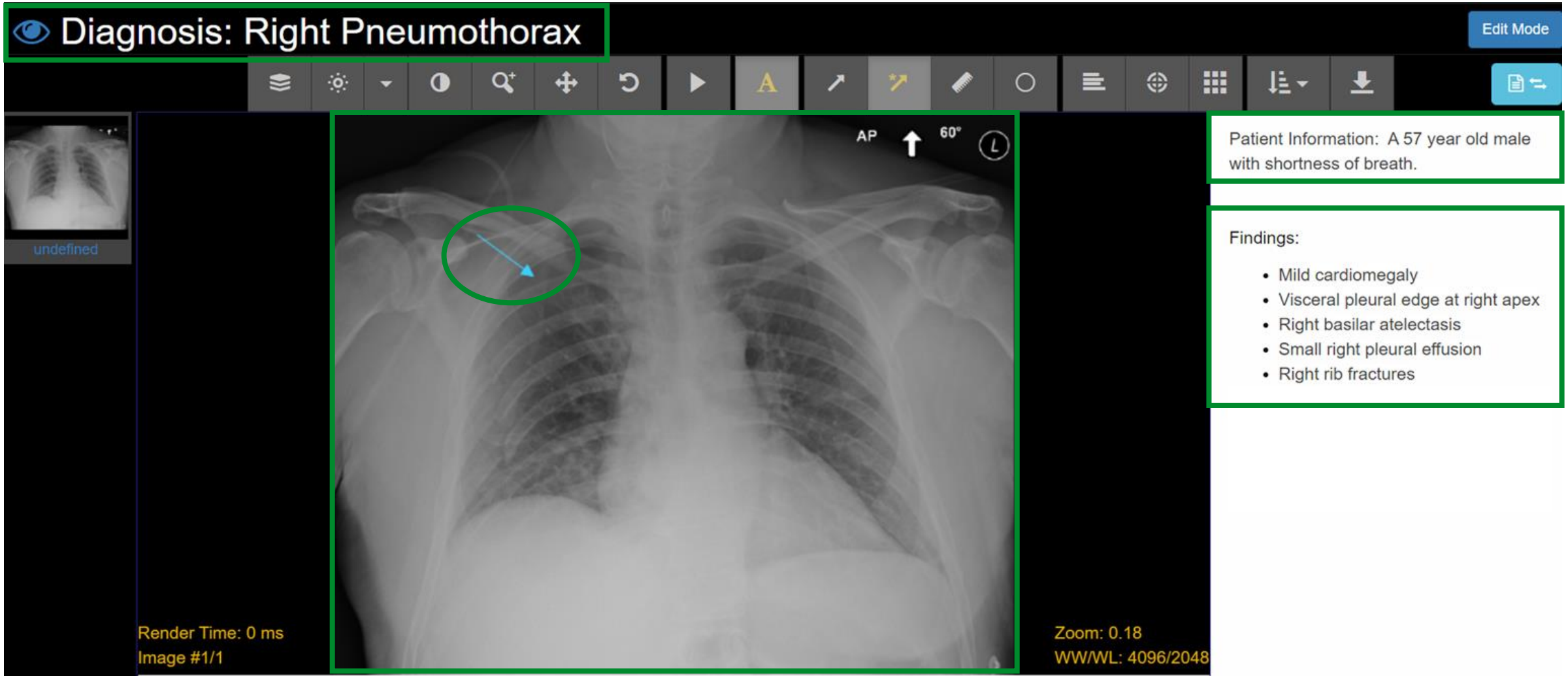
Übersicht
- Online-Experiment
- 138 Radiologen*innen
- 127 Internisten*innen/ Notfallmediziner*innen

- Design
  - Quelle der Empfehlung: Zwischensubjektfaktor
  - Grad der lokalen Erklärbarkeit: Innersubjektfaktor

Abhängige Variablen
- Performanz in %
- Bewertung der Empfehlungsqualität
  - Zustimmung
  - Nützlichkeit
  - Vertrauenswürdigkeit
  - Zukünftige Nutzung
- Vertrauen in die eigene Entscheidung

Kontrollvariablen
- Berufliche Identifikation
- Berufliche Autonomie
- Einstellung zu KI
- Wissen über KI
- Erfahrung in Jahren

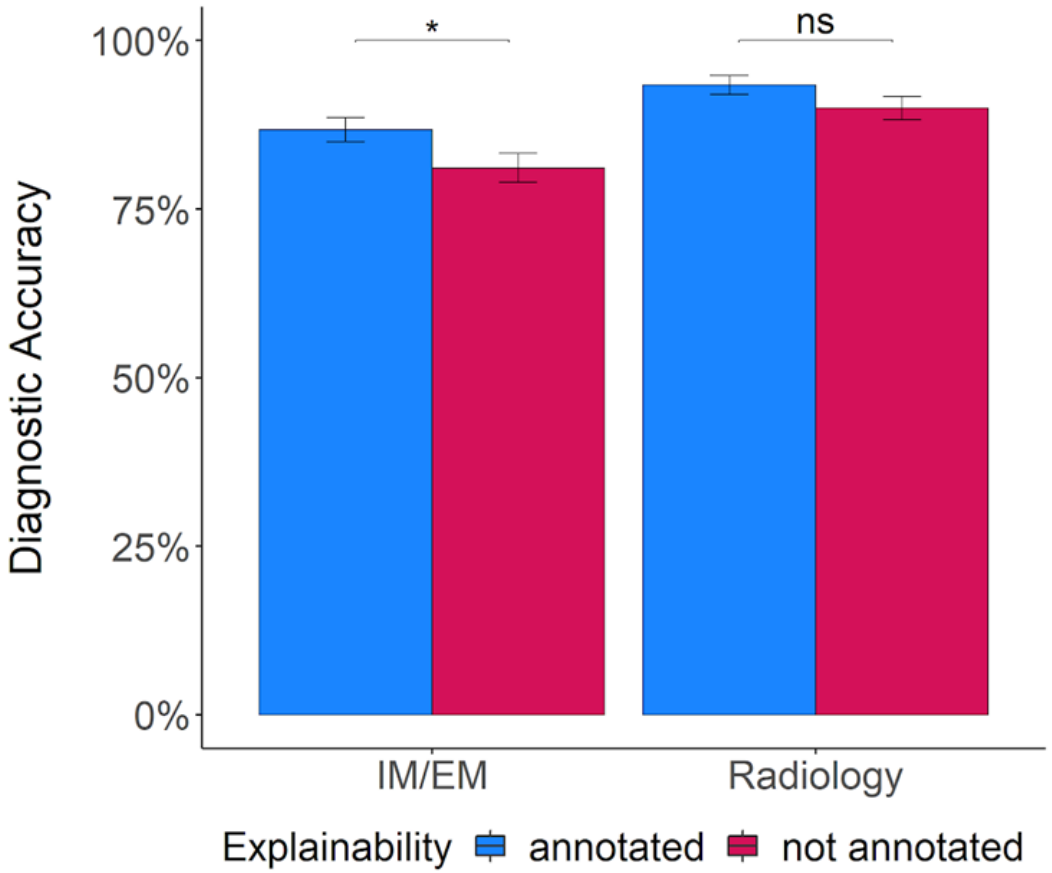
Methode Experiment 2

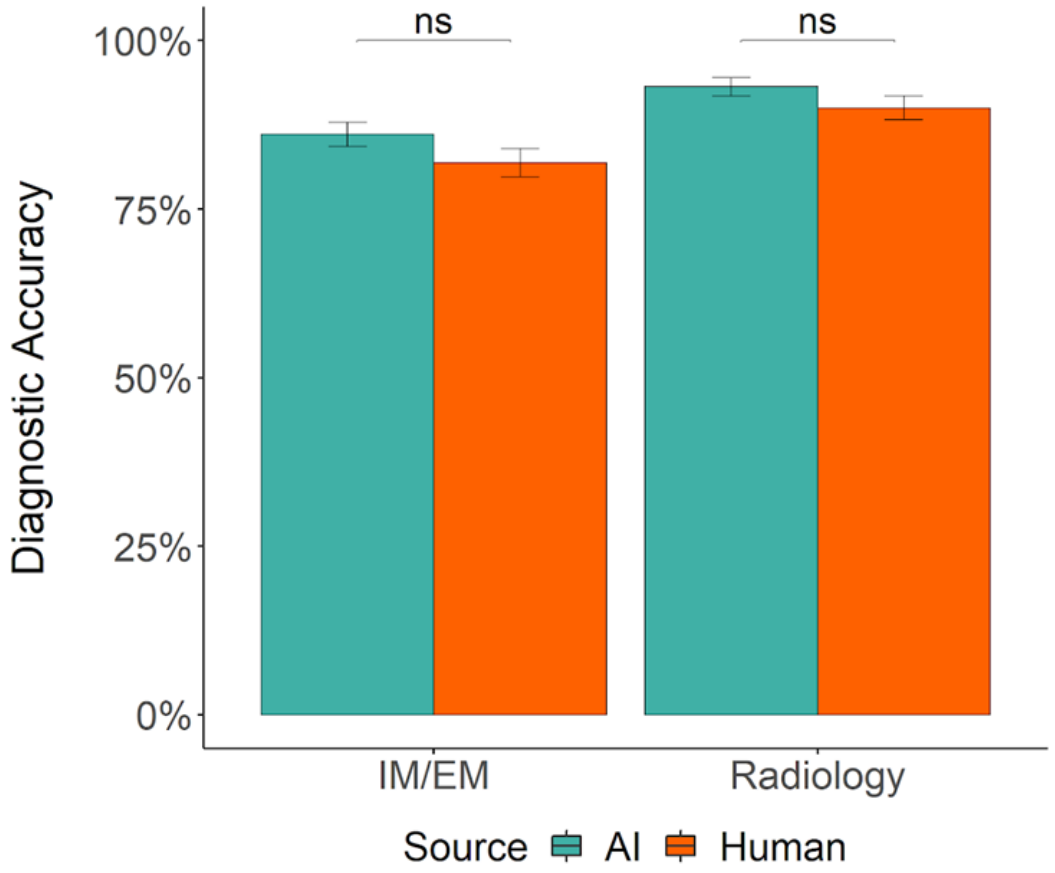


Logistic mixed multilevel regression models for participants' diagnostic accuracy.

Predictors	*Odds Ratios*	*SE*	*95% CI*	*z*	*p*
Intercept	1.30	1.38	0.16 – 10.37	0.24	0.807
Explainability [annotated]	2.30	0.61	1.37 – 3.86	3.17	**0.002**
Source [AI]	2.09	0.57	1.22 – 3.59	2.69	**0.007**
Task expertise [experts: radiologists]	2.20	0.51	1.40 – 3.47	3.41	**0.001**
Professional identification	1.00	0.12	0.80 – 1.26	0.02	0.985
Beliefs about professional autonomy	1.20	0.14	0.95 – 1.52	1.57	0.117
Self-reported AI-knowledge	1.34	0.19	1.01 – 1.77	2.03	**0.043**
Attitude toward AI	0.99	0.12	0.78 – 1.26	-0.06	0.954
Professional experience (years)	0.99	0.01	0.97 – 1.01	-1.11	0.268
Explainability [annotated] x Source [AI]	0.57	0.22	0.27 – 1.20	-1.49	0.137

Ergebnisse Experiment 2

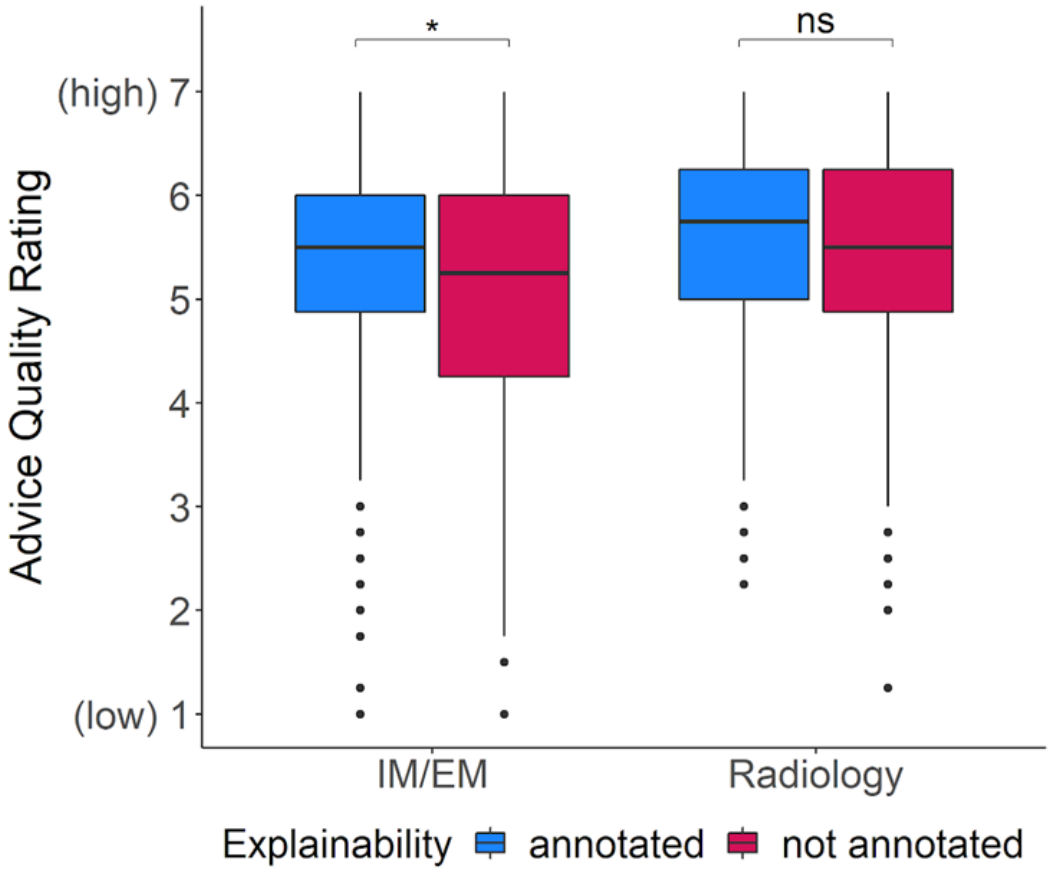


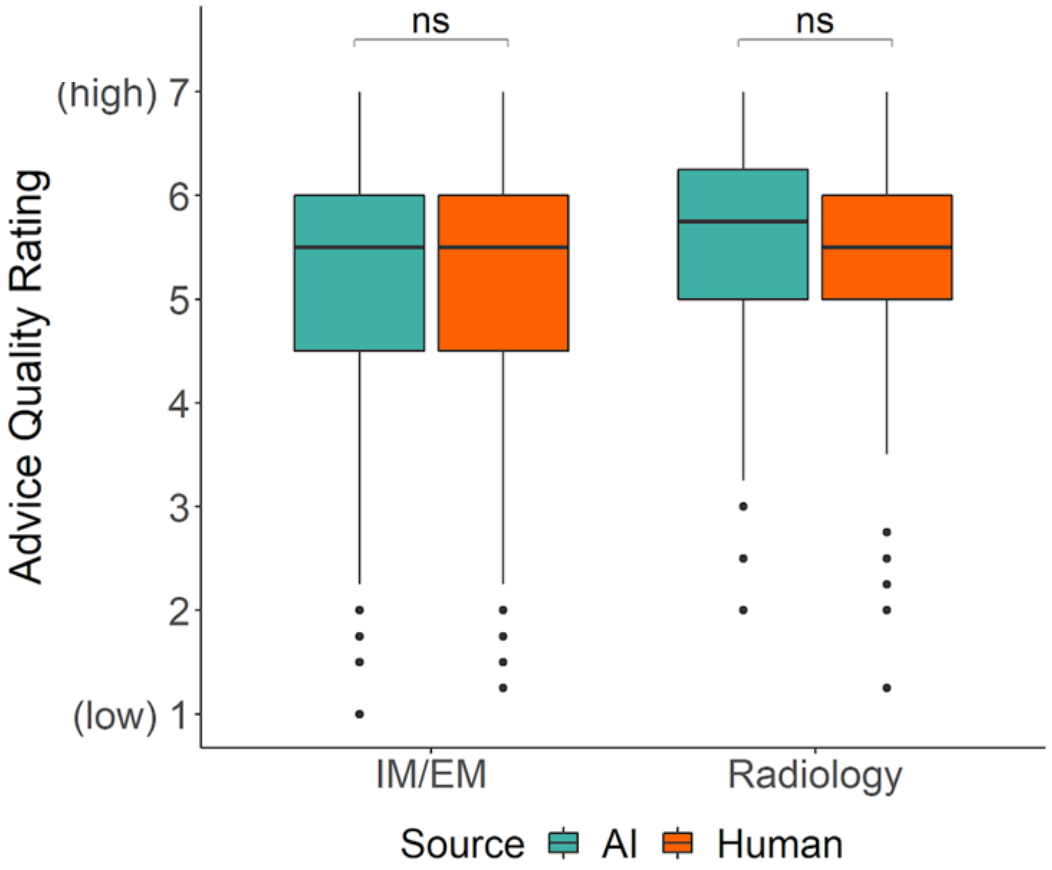


Linear mixed multilevel regression models for advice quality rating.

Predictors	*Estimate*	*SE*	*95% CI*	*t*	*p*
Intercept	4.55	0.49	3.59 – 5.50	9.36	**<0.001**
Explainability [annotated]	0.25	0.07	0.11 – 0.39	3.42	**0.001**
Source [AI]	0.03	0.11	-0.19 – 0.25	0.29	0.771
Task expertise [experts: radiologists]	0.24	0.10	0.04 – 0.45	2.32	**0.020**
Professional identification	0.03	0.05	-0.07 – 0.14	0.59	0.554
Beliefs about professional autonomy	-0.09	0.05	-0.20 – 0.01	-1.70	0.089
Self-reported AI-knowledge	0.07	0.07	-0.06 – 0.20	1.08	0.279
Attitude toward AI	0.13	0.06	0.02 – 0.23	2.27	**0.023**
Professional experience (years)	-0.00	0.01	-0.02 – 0.01	-0.90	0.367
Explainability [annotated] x Source [AI]	-0.07	0.10	-0.27 – 0.12	-0.75	0.454

Ergebnisse Experiment 2

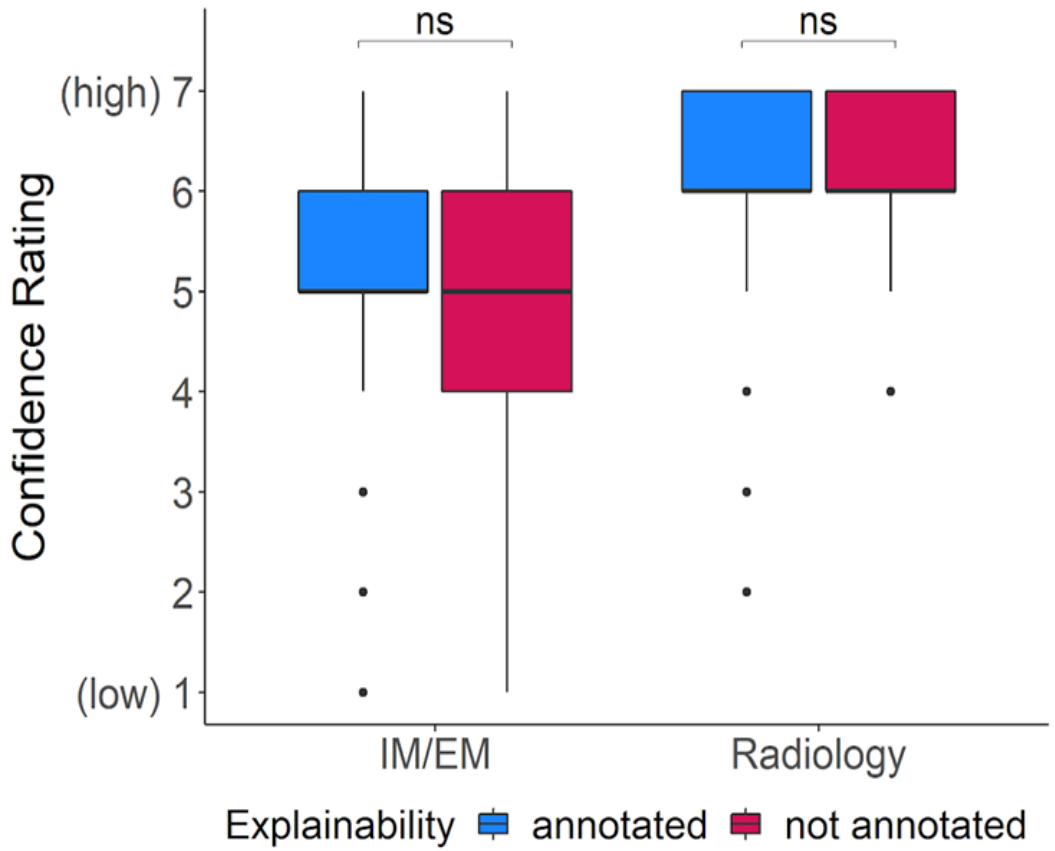


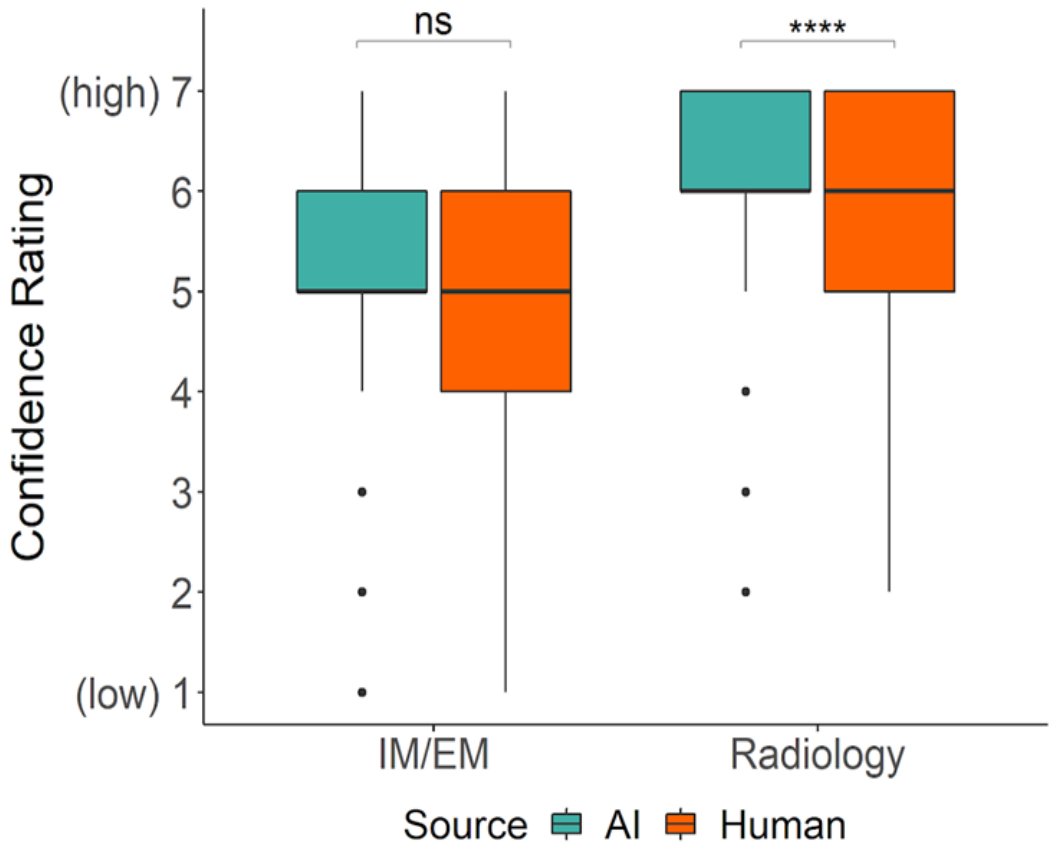


Linear mixed multilevel regression models for confidence in the diagnosis.

Predictors	*Estimate*	*SE*	*95% CI*	*t*	*p*
Intercept	4.60	0.43	3.76 – 5.44	10.76	**<0.001**
Explainability [annotated]	0.06	0.07	-0.09 – 0.20	0.77	0.440
Source [AI]	0.19	0.10	-0.01 – 0.39	1.91	0.056
Task expertise [experts: radiologists]	0.72	0.09	0.54 – 0.89	7.93	**<0.001**
Professional identification	-0.00	0.05	-0.09 – 0.09	-0.05	0.956
Beliefs about professional autonomy	0.01	0.05	-0.09 – 0.10	0.16	0.874
Self-reported AI-knowledge	0.19	0.06	0.08 – 0.30	3.39	**0.001**
Attitude toward AI	-0.01	0.05	-0.11 – 0.08	-0.26	0.794
Professional experience (years)	0.01	0.00	-0.00 – 0.02	1.79	0.074
Explainability [annotated] x Source [AI]	-0.03	0.10	-0.23 – 0.17	-0.32	0.749

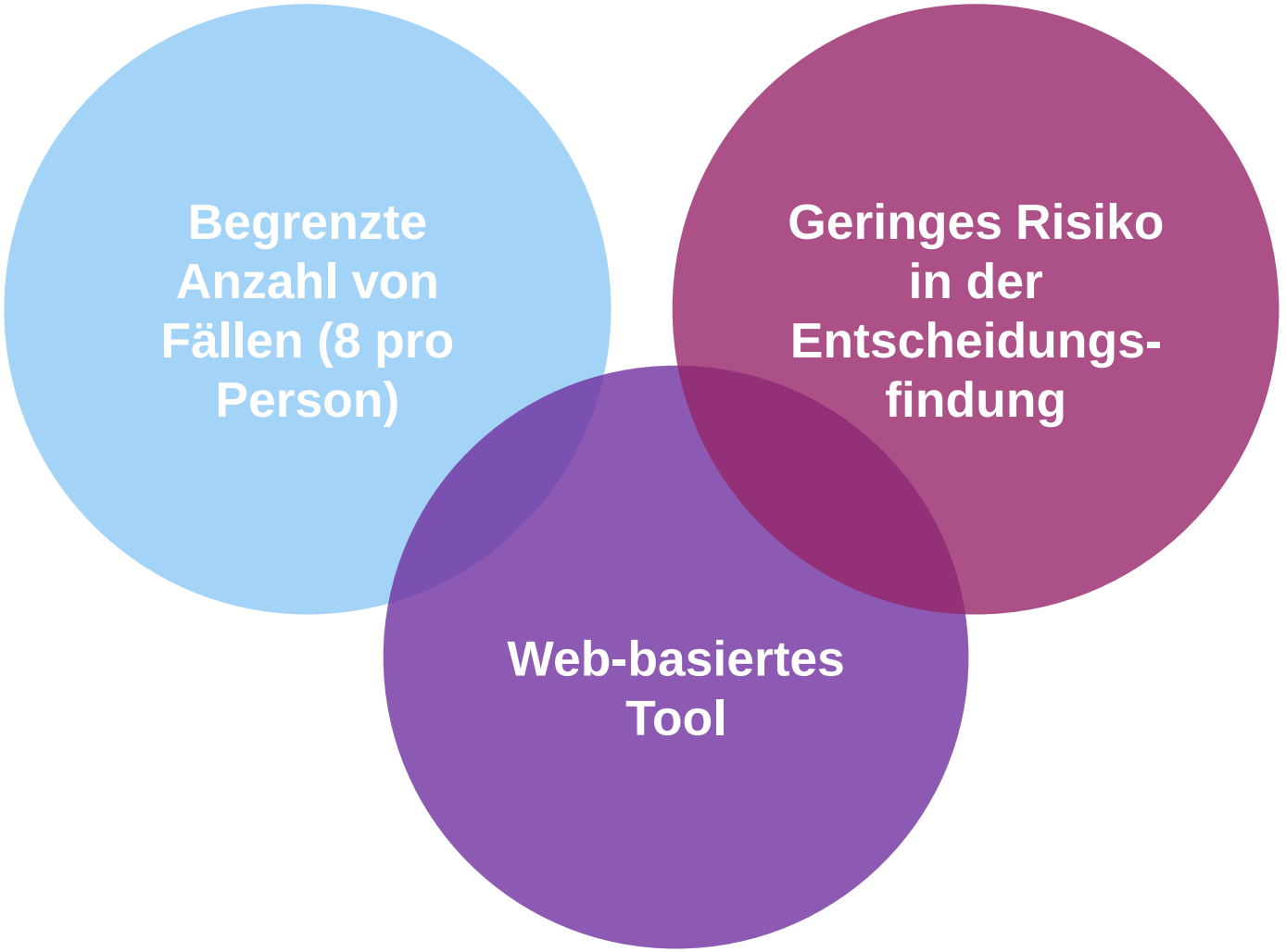
Ergebnisse Experiment 2



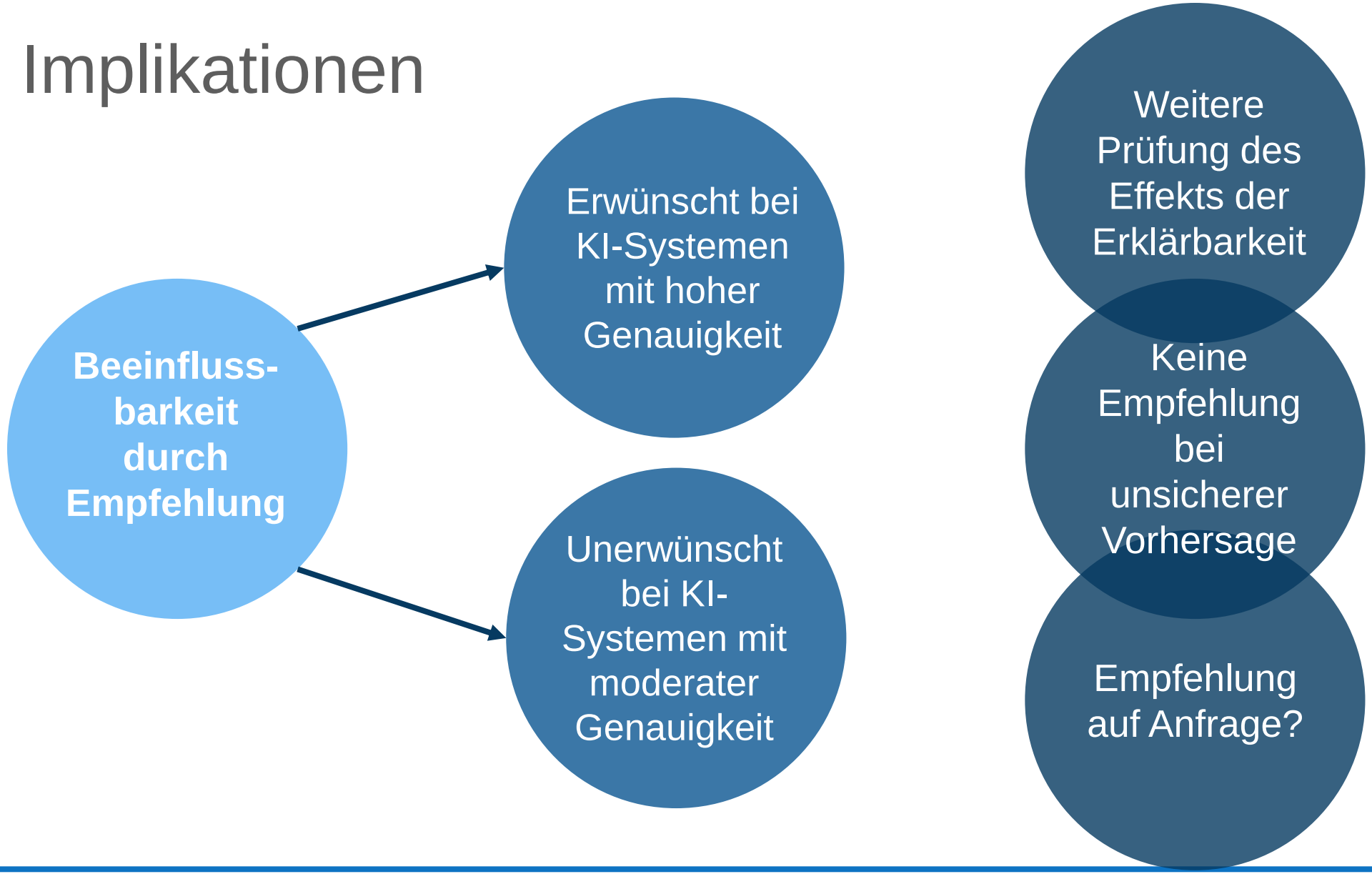


Diskussion: Experiment 2

Diskussion: Limitationen (Experiment 1 & 2)



Implikationen



VIELEN DANK FÜR IHRE AUFMERKSAMKEIT!

Harini Suresh
MIT

Martina Raue
MIT

Eva Lermer
LMU & HS Augsburg

Timo Koch
LMU

Marzyeh Ghassemi
MIT

Errol Colak
UoT

Dr. Susanne Gaube

LMU München & Universitätsklinikum Regensburg

susanne.gaube@psy.lmu.de
susanne.gaube@ukr.de

https://www.clinaid-lab.com
https://www.peoplemanagement.uni-muenchen.de

Funded by

VolkswagenStiftung