Universität St.Gallen
School of Medicine

krebsliga ostschweiz

Optimal utilization of an AI diagnostic software in a mammography screening program in Switzerland

Marcel Blum[1,2], Rudolf Morant[2], Alexander Geissler[1],
Justus Vogel[1], Jonas Subelack[1], David Ehlig[1]

1) University of St.Gallen, School of Medicine, Chair of Health Economics, Policy and Management, St.Gallen (CH)
2) Cancer League of Eastern Switzerland & "donna" Screening Program, St.Gallen (CH)

European Congress of Radiology 2025
RPS 705 - Artificial intelligence in breast imaging

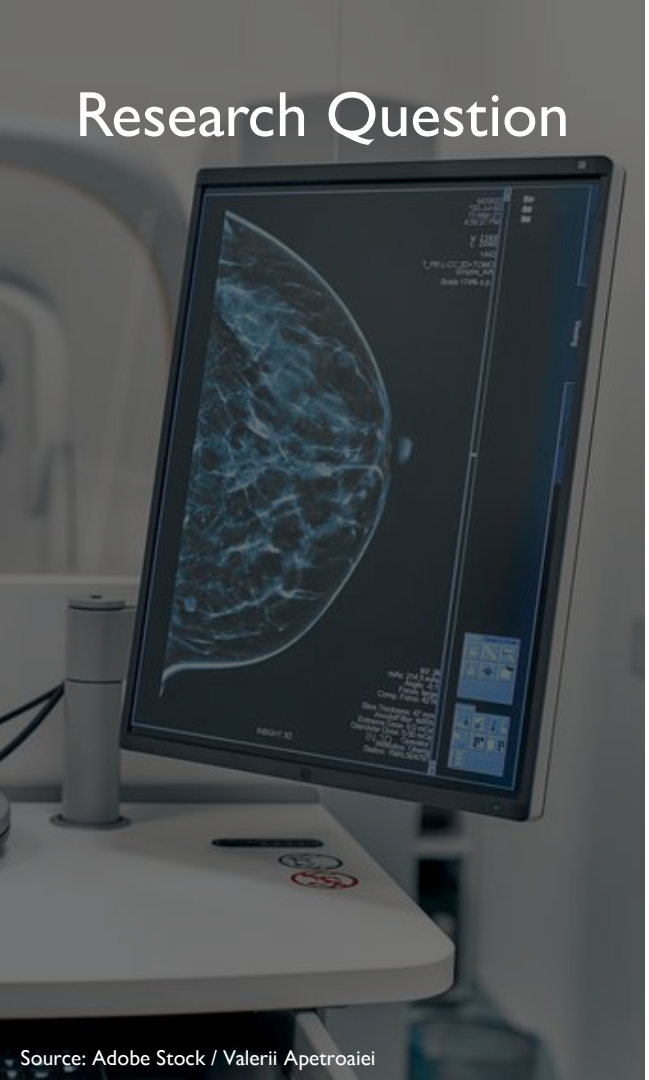
February 27th, 2025

From insight to impact.

Conflict of Interest Statement

- None of the authors have any conflict of interest to disclose.

Research Question



Source: Adobe Stock / Valerii Apetroaiei

- Artificial Intelligence (AI) has been tested and/or implemented in several mammography screening programs – yet its optimal utilization is not known
- AI is associated with a higher effectiveness (i.e., sensitivity and specificity) and/or a higher efficiency (i.e., workload)

- What is the optimal utilization of an AI diagnostic software in a mammography screening program in Switzerland?
- How would effectiveness and efficiency have been affected if AI had been used?

Data & Methods

"donna" Mammography Screening Program:

- Bi-annual invitations for 50- to 69/74-year-old women in six Swiss cantons
- ~ 54,000 mammographies per year (~ 50% participation rate)

Data of Interest: All mammographies of 50- to 69-year-old women from the canton of St.Gallen who participated in the "donna" program in 2022 and 2023

Data sources: Data linkage between the "donna" Screening Program, the Cancer Registry of Eastern Switzerland, and the AI scores provided by the ProFound AI® by iCAD, inc.

AI Case score for each 2D FFDM:

- Numeric from 0 to 100
- Reflects the degree of confidence that the case contains a cancer compared to the training database

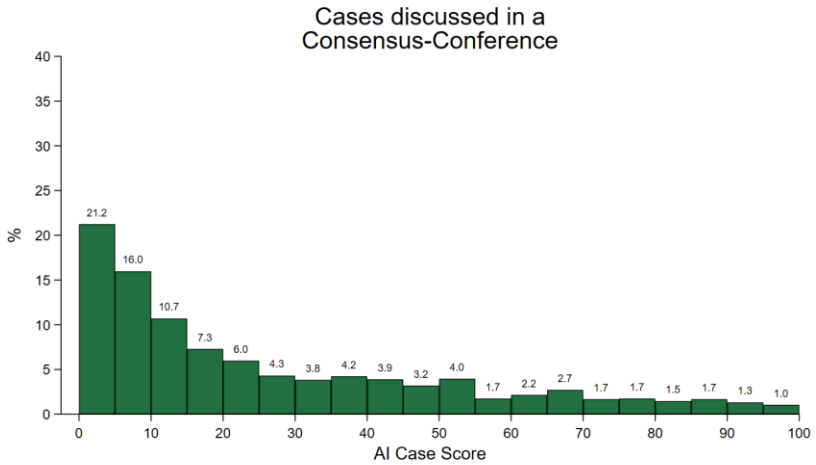
22,813 mammographies

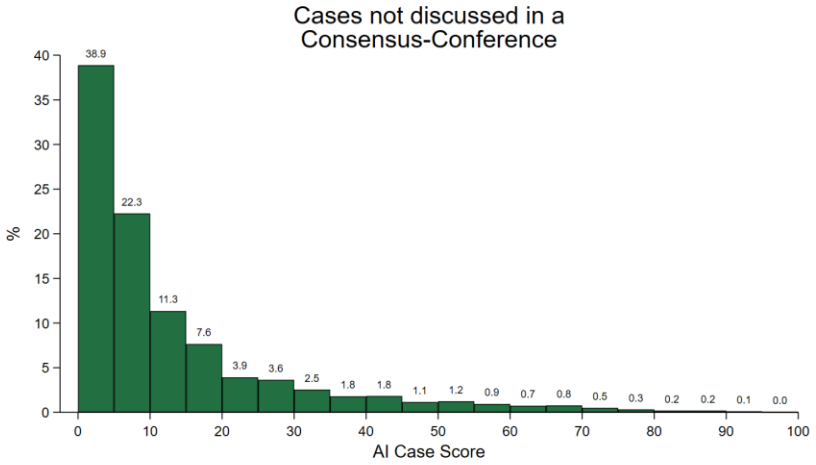
- **117 screen-detected breast carcinoma (SDC)**
- **35 interval breast carcinoma (IBC)***

* Preliminary number

Discordance between radiologists and AI

AI-Review whether case was initially discussed in a Consensus Conference (CC)



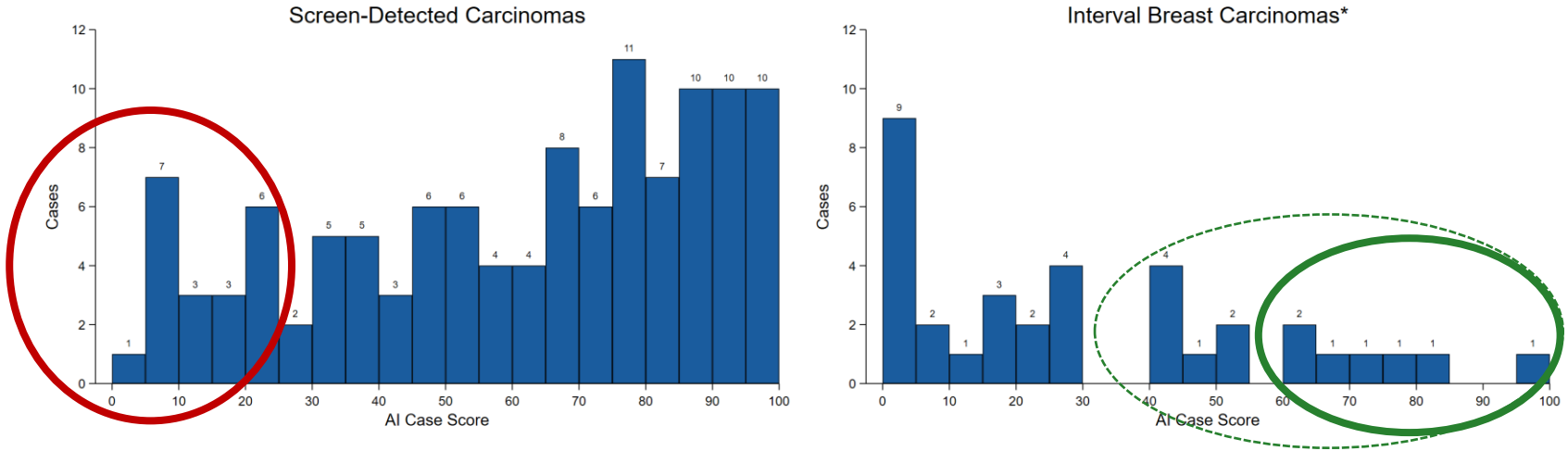


- Cases that were discussed in a CC had significantly higher case scores (Median=16 [IQR=6-42] vs. 7 [3-16], p<0.01)
- Of all cases with an AI score ≥ 48 (representing 6.3%, the same as the initial CC-rate) a fifth were discussed in a CC

➢ **The level of discordance between radiologists and AI seems to be higher than expected**

Need for AI-Radiologist collaboration

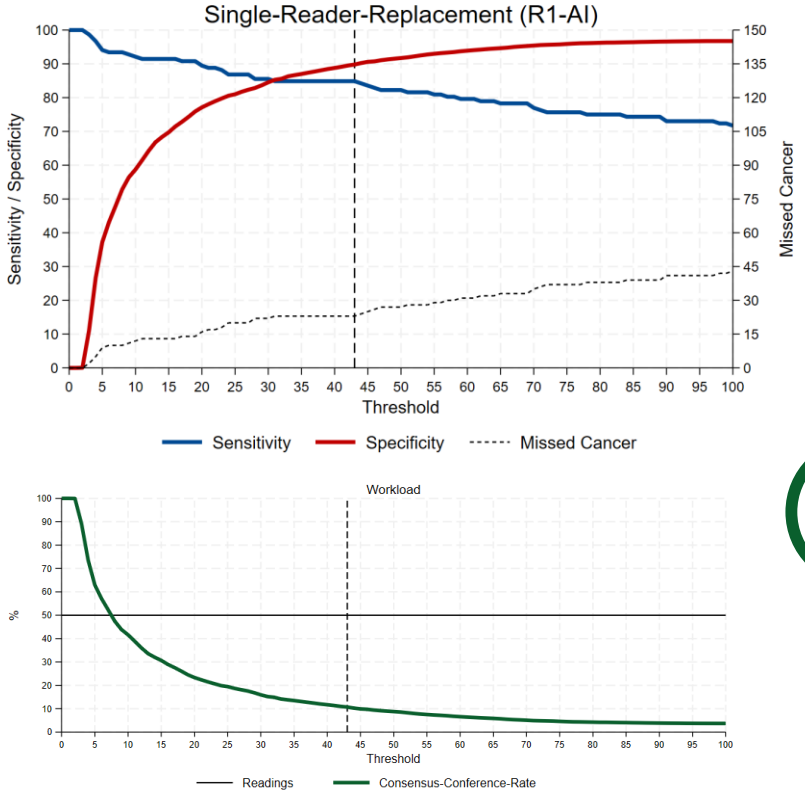
AI-Review of Cancer Cases (SDC or IBC)



- Half of the 117 SDCs had a case score ≥ 67, but some were AI-reviewed with an unsuspiciously low case score
- Conversely, AI might have detected a few IBCs, especially if they were not initially discussed in a CC

➢ **This emphasizes that AI and human reader(s) could effectively complement each other**

AI-Implementation-Scenario: Single-Reader-Replacement (R1-AI)



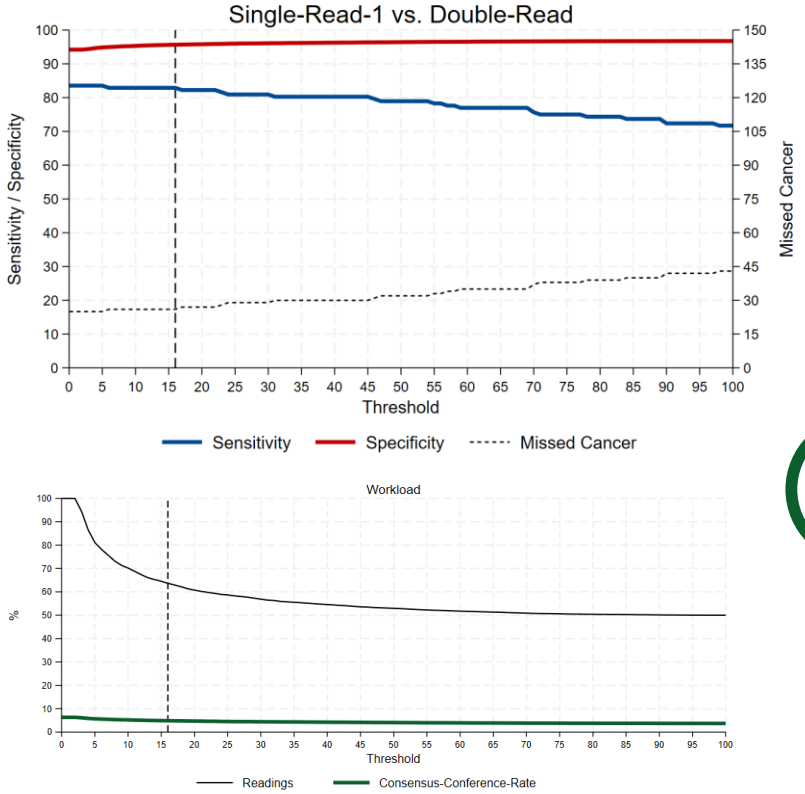
- Sensitivity and specificity are maximized according to the Youden Index at **threshold 43**, achieving a **higher sensitivity** (+1.3pp) but a **lower specificity** (-4.4pp) than Status-Quo (Double Read with CC)

- Substantial **workload reduction** (-46%) even if increase in CC-discussions (by 4.4pp to 10.7%)

Would have missed some (screen-detected) cancer cases but would have brought more IBCs cases to a CC

	Cancer	No Cancer
Case is discussed in a CC (either due to AI-Score above threshold or review by Reader 1)	TP	FP
Case is not discussed in a CC	FN	TN

AI-Implementation-Scenario: „Triage“ – Single-Read (R1) vs. Double-Read



- Sensitivity and specificity are maximized according to the Youden Index at **threshold 16**, achieving a **similar sensitivity** (-0.7pp) but a **higher specificity** (+1.5pp) than Status-Quo (Double Read with CC)
- Considerable **workload reduction** (-36%) and lower CC-discussion-rate (-1.4pp to 4.9%)

One SDC (along with 25 IBCs) would have been missed

	Cancer	No Cancer
Case is discussed in a CC	TP	FP
Case is not discussed in a CC	FN	TN

Conclusion

What is the optimal AI implementation in a Screening Program in terms of effectiveness and efficiency?

Case score

- Majority of cases are assigned with a low AI score (75% of cases < 17)
- Discordance between radiologists' and AI review

AI-Implementation

- Requirement of a human-AI interaction
- Single-Reader-Replacement or Triage are as effective as Status-Quo Double-Read, but with a ~40% workload reduction

These findings indicate a relevant potential of AI to safely reduce the workload of radiologists

Follow-up Research:
Full screening round (incl. all IBCs) as well as include further information on tumor characteristics and mammography devices

Thank you!

Marcel.Blum@unisg.ch