AI for reading screening mammograms: the need for circumspection

Philippe Autier¹ · Jean-Benoît Burrion² · André-Robert Grivegnée³

Received: 28 February 2020 / Accepted: 25 March 2020 / Published online: 21 April 2020
© European Society of Radiology 2020

Abstract

Key Points

- *The studies on AI reading of screening mammograms have methodological limitations that undermine the conclusion that AI could do better than radiologists.*
- *These studies do not inform numbers of extra breast cancers found by AI that could represent overdiagnosis.*
- *The ability of AI to detect breast cancers is overestimated because there are no result on biopsy procedures that should be performed when mammograms are positive at AI reading but not at radiology reading.*

The article of McKinney and colleagues claims that an artificial intelligence (AI) system outperforms radiologists for the identification of breast cancer in screening mammograms [1]. However, the study design has limitations and the way results are reported tends to obscure key data which do not support claims made by authors.

The study has included breast cancers detected by screening mammography and breast cancers diagnosed during a 3- (in the UK) or 2-year (in the USA) follow-up after screening. Hence, all cancers were those found by screening mammography plus the interval cancers. The cancers found by screening mammography are a mix of tumours that would have progressed into clinical cancer in the absence of screening mammography, and of overdiagnosed tumours that would have not progressed into clinical cancer during women's lifetime. The interval breast cancers are a mix of cancers missed by screening mammography plus cancers that arose during follow-up.

Using results displayed in Extended Data Tables 2 and 5 of the publication, we could classify the 402 breast cancers of the UK part of the study in four mutually exclusive categories

(Table 1). Such reworking was not possible for mammograms that were negative at both readings. So, 252 cancers (63%) were found by the first radiologist, 263 (65%) were found by AI, and the two modalities were discordant for 20% of mammograms in which a cancer was present. Of note, 26% of cancers still arose as interval cancers during follow-up.

But the main limitation of the study is bound to its retrospective design for AI readings. Work-up of positive mammogram readings was done for radiologist readings only. There was no work-up of mammograms negative at radiologist reading but positive at AI reading. So, if no interval cancer showed up after a positive AI reading but negative radiologist reading, these AI readings were considered as false-positive results. However, work-up of mammograms positive at AI reading only could have found in situ and small, localised cancers that would not have progressed into clinical breast cancer, and thus not shown up as interval cancer. Hence, the study cannot inform on the amount of extra cancers that would be associated with AI reading. Had the study adopted a prospective design with work-up of all positive mammogram readings by radiologists or AI, there would be more than 45 cancers in the radiologist negative-AI positive category, the actual number of in situ cancers detected by AI would be greater than the 7 reported, and the total of breast cancers would be greater than 402.

The study suggests that 45 interval cancers would have been prevented thanks to AI. This is probably an overestimation. The sensitivity of biopsy procedures is not 100% [2, 3]. Therefore, a fraction of biopsies done for positive mammograms at AI reading only would have returned negative, in which case less than 45 interval cancers would have been averted thanks to AI and there would be more than 105 false-negative results for both readings.

✉ Philippe Autier
Philippe.autier@i-pri.org; <http://www.i-pri.org>

¹ University of Strathclyde Institute of Global Public Health, iPRI International Prevention Research Institute, Le Campus, Bâtiment L'Australien, 18 Chemin des Cuers, 69570 Dardilly, France

² Prevention and Screening Clinic, Jules Bordet Institute, Brussels, Belgium

³ Breast Radiology Unit, Jules Bordet Institute, Brussels, Belgium

Table 1 Breast cancers diagnosed following screening mammograms read by the first radiologist or using an AI system in the UK centres (McKinney et al 2020)

		AI readings		
		Positive	Negative	Total
1st radiologist readings	Positive	218 (54%)	34 (9%)	252
	Negative	45 (11%)	105 (26%)	150
	Total	263	139	402

The 52% false-negative rate in the USA part of the study is far greater than for the first reader in the UK, and greater than false-negative rate benchmarks cited for digital screening mammography in the USA [4]. Hence, the greater sensitivity and specificity of AI readings than radiologist readings in the USA could thus be due to a sub-optimal ability of radiologists to detect cancers in mammograms.

In conclusion, the McKinney et al study is not a valid comparison between radiologist readings and AI readings of mammograms. The few other large studies that compared radiologist reading to AI reading had similar limitations [5]. Various scenarios for the incorporation of AI systems in breast screening programmes are envisioned, from selection of mammograms to be read by radiologists (workload reduction), second reading purposes (minimisation of false negatives) and substitute of radiologists (e.g., in areas where radiologists are rare) [1, 6]. However, prospective studies embedded within routine screening practice are needed for getting a more reliable picture of the added value of AI systems, especially a better appraisal of the overdiagnosis that could be generated by AI readings, as well as a more accurate estimation of cancers detected when mammograms are positive on AI reading only.

Funding information The authors state that this work has not received any funding.

Compliance with ethical standards

Guarantor The scientific guarantor of this publication is Philippe Autier.

Conflict of interest The authors of this manuscript declare no relationships with any companies whose products or services may be related to the subject matter of the article.

Statistics and biometry One of the authors, Philippe Autier, has significant statistical expertise.

Informed consent Written informed consent was not required for this study because based on published data.

Ethical approval Institutional Review Board approval was not required because no human subjects or animals involved.

Methodology

- Not applicable.

References

1. McKinney SM, Sieniek M, Godbole V et al (2020) International evaluation of an AI system for breast cancer screening. *Nature* 577: 89–94
2. Georgieva RD, Obdeijn IM, Jager A, Hooning MJ, Tilanus-Linthorst MM, van Deurzen CH (2013) Breast fine-needle aspiration cytology performance in the high-risk screening population: a study of BRCA1/BRCA2 mutation carriers. *Cancer Cytopathol* 121:561–567
3. Wesola M, Jelen M (2013) The diagnostic efficiency of fine needle aspiration biopsy in breast cancers - review. *Adv Clin Exp Med* 22: 887–892
4. Lehman CD, Arao RF, Sprague BL et al (2017) National performance benchmarks for modern screening digital mammography: update from the breast cancer surveillance consortium. *Radiology* 283: 49–58
5. Rodriguez-Ruiz A, Lang K, Gubern-Merida A et al (2019) Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. *J Natl Cancer Inst* 111: 916–922
6. Rodriguez-Ruiz A, Lang K, Gubern-Merida A et al (2019) Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? A feasibility study. *Eur Radiol* 29:4825–4832

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.