Title

Artificial intelligence and magnetic resonance imaging may not make cancer screening better.

Authors

Powell, Kerrington
Kim, Myung S
Haslam, Alyson
et al.

Publication Date
2022-03-01

DOI
10.1016/j.jcpo.2021.100314

Journal of Cancer Policy 31 (2022) 100314

Contents lists available at ScienceDirect

Journal of Cancer Policy

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

Artificial intelligence and magnetic resonance imaging may not make cancer screening better

1. Introduction

While there is great potential for imaging and artificial intelligence to improve cancer screening, two recent studies suggest pitfalls and challenges. First, researchers at Google applied artificial intelligence (AI) to mammographic images to improve breast cancer screening [1]. Next, Eklund et al. [2] recently show that magnetic resonance imaging (MRI) guided biopsy for prostate cancer screening may improve the diagnosis of clinically significant lesions over the current standard of blind biopsy. Both papers claim a novel approach to cancer screening is less likely to identify pre-malignant, indolent or low concern lesions, and more likely to identify high grade or high concern lesions. However, neither study employs a robust conceptual framework. Here, we provide a framework to think about cancer screening, and argue that histopathological findings—the gold standard of both these studies—are inadequate substitutes for what screening seeks to find.

2. Google AI paper

In a recent study published by Google Health, researchers tested an AI system for breast cancer screening using a retrospective dataset containing mammograms from the United States (USA, Northwestern Memorial Hospital) and the United Kingdom (UK, multiple screening locations) [1]. In the clinical comparison evaluating mammograms from the USA and UK datasets, the researchers discovered that the AI system outperformed human readers in terms of absolute sensitivity and specificity. Specifically, the AI sensitivity decision was 65.4 % and 57.5 % whereas the clinical decision was 62.7 % and 48.1 % in the UK and USA datasets, respectively. Corresponding and in relation to absolute specificity, the AI decision was 94.1 % and 86.5 % while the clinical decision was 92.9 % and 80.8 %. In a head-to-head comparison against radiologists, AI identified additional cancers otherwise missed, and these cancers were more likely to be invasive rather than confined. At first glance, AI seems to outperform and improve areas such as efficiency, accuracy, detection, and workload [1].

3. MRI guided prostate cancer screening

The randomized controlled trial (STHLM3-MRI, NCT03377881) demonstrated that MRI-targeted biopsy was non-inferior compared with standard biopsy for the detection of clinically significant (Gleason score $\geq$7) prostate cancer in males aged 50–74 with prostate-specific antigen (PSA) levels $\geq$3 ng/mL [2]. Specifically, 21 % of men were diagnosed in the experimental arm compared to 18 % in the control arm, falling within the non-inferiority margin of 4 percentage points. Additionally, MRI-targeted biopsy was reported to detect fewer clinically insignificant cancers (4 % vs. 12 %, Gleason score $\leq$ 6) and benign findings [2]. Readers may conclude from the trial's report that MRI guided biopsy preserves detection of concerning cancers, and minimizes detection of benign and insignificant lesions.

4. The problem

Cancer screening aims to find tumors earlier, when surgery is possible, and thereby prevent cancer deaths. However, screening is complicated by cancer's diverse mutational profiles, meaning progression is difficult to predict. Because of this high degree of heterogeneity, a "barnyard" analogy was devised in an attempt to distill these concepts [3]. The comparison makes use of (1) turtles to represent indolent cancers that will not cause harm or death during an individual's natural life span; (2) rabbits to symbolize malignancies that are destined to progress but have not yet metastasized and may be treated with surgery or radiation; (3) and birds to depict aggressive tumors that have already metastasized and will be lethal no matter when they are found (Fig. 1) [3]. The objective of screening is to find more rabbits, people who would likely benefit from screening, while limiting indolent and aggressive disease detection. Which brings us to both papers: did these screening techniques find more rabbits?

5. The core challenges

Both studies suffer from the inability to differentiate between rabbits, birds and turtles [4]. In other words, although these novel methods discover more significant lesions and fewer low-risk lesions based on histopathology; those histopathologic findings do not directly translate into the three cancer categories (birds, rabbits & turtles). This is shown in the Figure. There is often large overlap between traditional measures on biopsy, and these 3 groups.

Put differently, even if novel screening methods outperform human readers, they may be counterproductive to the goal of screening if they find more slowly proliferating oddities and malignancies destined for rapid progression, and fewer cancers treatable with early detection. The use of invasion or Gleason is no guarantee you are finding what we seek (rabbits); Because at each Gleason level, or among invasive ductal cancer, you have all three categories.

Second, pathologic staging is done for prognostic reasons for various cancer types, however, the grading systems may not always accurately reflect cancer's propensity to grow aggressively or not. In breast cancer, diagnosis of ductal carcinoma in situ (DCIS) is an example of this issue [3]. Since no one can predict whether DCIS will progress or not, potentially leading to overtreatment and unnecessary psychological stress.

Third, in prostate cancer, a Gleason score from an MRI-targeted biopsy and a Gleason score from a standard biopsy may have different

https://doi.org/10.1016/j.jcpo.2021.100314

Available online 27 November 2021

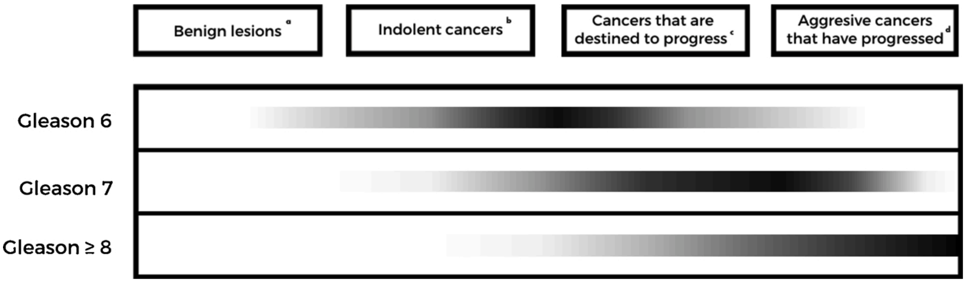
Benign lesions [a]
Indolent cancers [b]
Cancers that are destined to progress [c]
Aggresive cancers that have progressed [d]
Gleason 6
Gleason 7
Gleason ≥ 8

Fig. 1. [a] Non-cancerous lesion. [b] Indolent cancers that will not cause harm or death during an individual's natural life span [Turtle]. [c] Malignancies that are destined to progress but have not yet metastasized [Rabbit]. [d] Aggressive tumors that have already metastasized and may be lethal [Bird].

prognostic implications. For example, a Gleason grade group 3 (Gleason score of 4 + 3 [7]) has the same score as a Gleason grade group 2 (Gleason score of 3 + 4 [7]) in different proportions, yet has a three-fold increase in prostate cancer mortality when obtained from a standard biopsy [5]. However, when MRI-targeted biopsies are considered, the same tumor may result in a Gleason score of 4 + 3 in an MRI-guided biopsy and a Gleason score of 3 + 4 in a conventional biopsy due to better targeting of high-grade lesions using direct visualization scores, such as the Prostate Imaging Reporting and Data System (PI-RADS) [2]. Simply put, the Gleason grade may be artificially increased for individuals who would otherwise receive a lower grade in a random sample, subjecting them to unnecessary treatment. This would imply that the Gleason grade group from an MRI sample is not representative because: (1) Pathologic guidelines recommend that the highest grade core from among the many cores obtained through an MRI-targeted biopsy be utilized; (2) When a biopsy is obtained in a novel manner, it is necessary to re-establish the relationship's strength; failure to do so implies that cancers of the same grade harbor the same risk, regardless of the biopsy method [6]. Therefore, prognostic associations and management recommendations cannot be reliably extrapolated to the MRI-targeted biopsy because the former relationships were derived based on standard biopsy data.

Fourth, although the Gleason score is a surrogate marker for effective cancer screening, certain high-grade tumors may already have undetectable distant metastases, in which case early detection has limited benefit. Also, research on screening tests indicate that individuals with negative MRIs may nevertheless have clinically significant prostate cancer [7]. In the NEJM trial, patients with negative MRIs did not get a biopsy [2]. This will certainly reduce benign lesion biopsies, but with the marginal benefits of traditional prostate cancer screening, missing just a few "rabbits" may tilt the risk-benefit balance in the wrong way. Furthermore, standard biopsies detected additional clinically significant prostate cancer that MRI-guided biopsies missed, and by removing the supplement standard biopsy included in the MRI-targeted biopsy procedure, the intervention no longer satisfies the non-inferiority margin (i.e., fewer clinically significant cancers are identified) [2]. This implies that if MRI imaging were so transformational, no further biopsy would be needed to preserve non-inferiority for clinically significant cancers.

6. What can be done?

Because there is no definitive feature on biopsy that can differentiate rabbits from turtles from birds, research on enhanced screening methods must directly measure survival and quality of life measures. In the Google AI study, the ratio with which the barnyard tumors are detected determines whether or not women benefit from the modality [1]. In the MRI-targeted therapy study, only long-term studies showing that MRI guided biopsy does not result in worse prostate cancer outcomes and does so with fewer biopsy and downstream procedures can validate the technique.

Due to these constraints, we suggest that the next MRI-targeted research protocol include three arms: (1) No screening; (2) Annual PSA screening + standard biopsy; (3) and Annual PSA screening + MRI-targeted biopsies. We do not yet know if prostate cancer screening improves overall survival or quality of life, justifying a no-screening control arm. For breast cancer, a randomized trial in accordance with current breast cancer screening standards, as well as the inclusion of three arms, are suggested for the AI system: (1) Conventional mammography; (2) Mammography using just AI; (3) Mammography with AI assistance. These trials should have adequate power to detect the proposed primary endpoint, cancer mortality, as well as important secondary endpoints, such as all-cause mortality and quality of life. We are still unaware of any physical feature that distinguishes the cancers we want to discover from those we don't want to find, and in the lack of empiricism, we must rely on randomized data. Whether or not these novel methods lead to clinical benefit in patients is a hypothesis that remains uninterrogated, and to best serve our patients, we must demonstrate our commitment by pursuing these answers.

Authorship contribution

VP & AH conceptualized study design; KP and MK reviewed literature; VP, AH and MK reviewed and confirmed abstracted data; KP wrote the first draft of the manuscript; and all authors reviewed and revised subsequent and finalized draft of the manuscript.

Funding

This study was funded by Arnold Ventures.

Declaration of Competing Interest

Vinay Prasad's Disclosures. (Research funding) Arnold Ventures (Royalties) Johns Hopkins Press, Medscape, MedPage (Consulting) UnitedHealthcare. (Speaking fees) Evicore. New Century Health (Other) Plenary Session podcast has Patreon backers. All other authors have no financial nor non-financial conflicts of interest to report.

References

[1] S.M. McKinney, M. Sieniek, V. Godbole, et al., International evaluation of an AI system for breast cancer screening, Nature 577 (2020) 89–94.

[2] M. Eklund, F. Jäderling, A. Discacciati, et al., MRI-targeted or standard biopsy in prostate cancer screening, N. Engl. J. Med. (2021), https://doi.org/10.1056/nejmoa2100852.

[3] H.G. Welch, The heterogeneity of cancer, Breast Cancer Res. Treat. 169 (2018) 207–208, https://doi.org/10.1007/s10549-018-4691-4.

[4] A.S. Adamson, H.G. Welch, Machine learning and the cancer-diagnosis problem - no gold standard, N. Engl. J. Med. 381 (24) (2019) 2285–2287, https://doi.org/10.1056/NEJMp1907407.

[5] M.S. Pepe, T.A. Alonzo, Comparing disease screening tests when true disease status is ascertained only for screen positives, Biostatistics 2 (2001) 249–260.

[6] A.J. Vickers, Effects of magnetic resonance imaging targeting on overdiagnosis and overtreatment of prostate cancer, Eur. Urol. (2021), https://doi.org/10.1016/j.eururo.2021.06.026.

[7] J.R. Stark, S. Perner, M.J. Stampfer, et al., Gleason score and lethal prostate cancer: does 3 + 4 = 4 + 3? J. Clin. Oncol. 27 (21) (2009) 3459–3464, https://doi.org/10.1200/JCO.2008.20.4669.

Kerrington Powell
College of Medicine, Texas A&M Health Science Center, Bryan, TX, 77807, USA

Myung S. Kim
Division of Hematology and Medical Oncology, Knight Cancer Institute, Oregon Health and Science University, Portland, USA

Alyson Haslam, Vinay Prasad*
University of California San Francisco, 550 16th St, 2nd Fl, San Francisco, CA, 94158, USA

* Corresponding author at: Department of Epidemiology and Biostatistics, University of California San Francisco, 550 16th St, 2nd Fl, San Francisco, CA, 94158, USA.
E-mail address: vinayak.prasad@ucsf.edu (V. Prasad).