verily

Clinical Evidence Generation and Integrated Diagnostics

Incorporating Integrated Diagnostics into Precision Oncology Care National Academies Workshop

Brad Hirsch, MD
Head of Product and Implementation, Verily

February 2023

Disclosures

I am an employee of Verily and serve on the Board of Directors of Galvani Bioelectronics.

Our origin story

2013

Born in the Google [X] moonshot factory

Google

2016

Founded as Verily Life Sciences

Alphabet

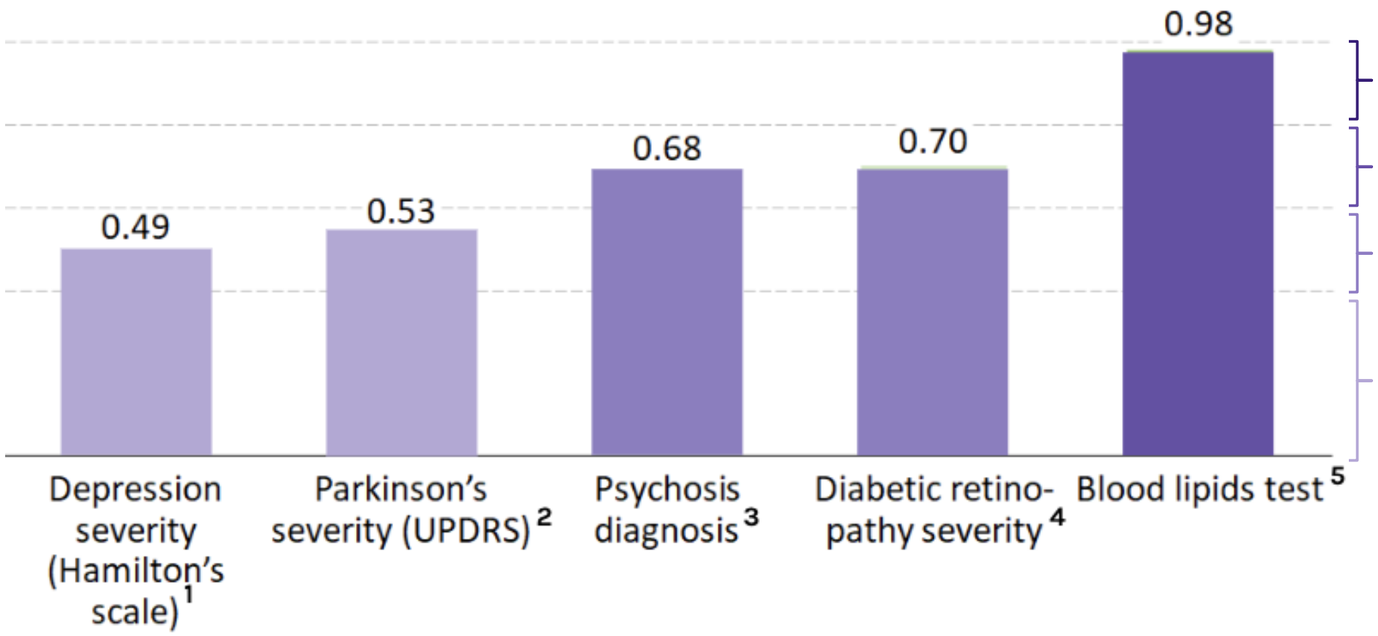
Today

A precision health company focused on finding the best health solutions for each person & population

- **1,500** employees
- **1,013** patents & applications
- **Headquartered in the U.S.**, offices in Canada, United Kingdom, Ireland, Israel and Singapore

Subjectivity reduces accuracy of key disease metrics

Inter-rater agreement (kappa) scores for common clinical trial endpoints



Quality of agreement[1]

- **Great** *(0.8-1.0)*
- **Good** *(0.6-0.8)*
- **Fair** *(0.4-0.6)*
- **Poor** *(0.0-0.4)*

1 Morriss et al (2008), **2** Palmer et al (2011), **3** Einfeld et al (2008), **4** Thapa et al (2016); **5** Bastianelli (2017)

Screening for diabetic retinopathy

THE LANCET
Digital Health

Real-time diabetic retinopathy screening by deep learning in a multisite national screening programme: a prospective interventional cohort study

Paisan Ruamviboonsuk, Richa Tiwari, Rory Sayres, Variya Nganthavee, Kornwipa Hemarat, Apinpat Kongprayoon, Rajiv Raman, Brian Levinstein, Yun Liu, Mike Schaekermann, Roy Lee, Sunny Virmani, Kasumi Widner, John Chambers, Fred Hersch, Lily Peng, Dale R Webster

Background

Diabetic retinopathy is a leading cause of preventable blindness, especially in low-income and middle income countries (LMICs). Deep-learning systems have the potential to enhance diabetic retinopathy screenings in these settings, yet prospective studies assessing their usability and performance are scarce.

Source: The Lancet

Screening for diabetic retinopathy

Expanding access to quality eye screening

Verily Retinal Service aims to reduce preventable blindness by empowering providers with technology that enables the detection of diabetic retinopathy (DR) by clinicians.

LEARN MORE

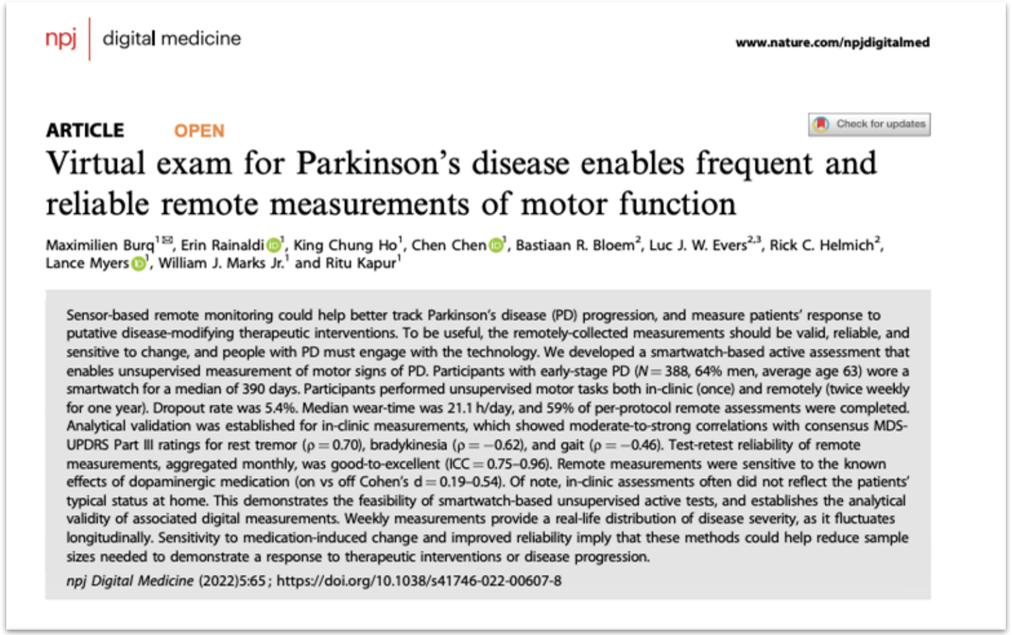
Image credit: Verily

Digital biomarkers in Parkinson's disease

We have built **a platform that uses wearable sensors for remote clinical assessments**.

The first instance is the Virtual Motor Exam for Parkinson's disease.

Image credits: Verily
Source: npj Digital Medicine



npj | digital medicine

www.nature.com/npjdigitalmed

ARTICLE OPEN

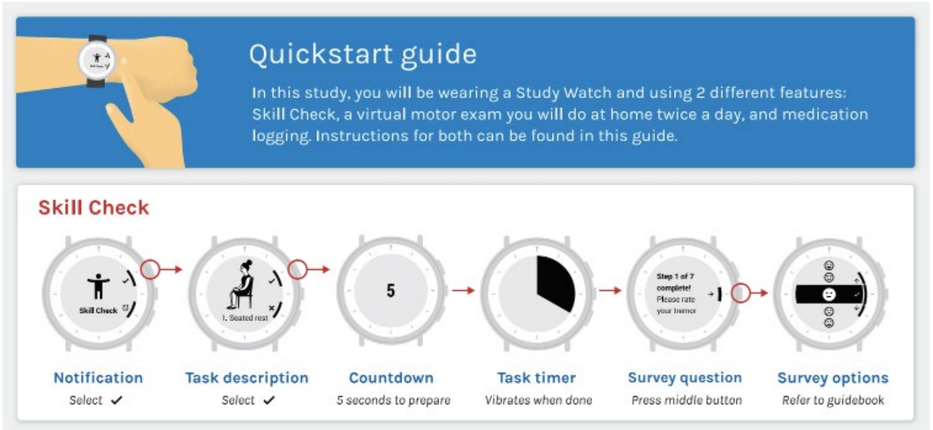
Check for updates

Virtual exam for Parkinson's disease enables frequent and reliable remote measurements of motor function

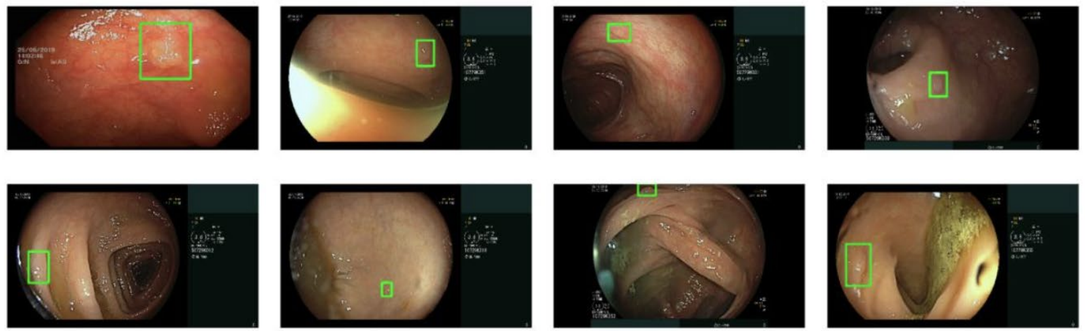
Maximilien Burq[1✉], Erin Rainaldi[1], King Chung Ho[1], Chen Chen[1], Bastiaan R. Bloem[2], Luc J. W. Evers[2,3], Rick C. Helmich[2], Lance Myers[1], William J. Marks Jr.[1] and Ritu Kapur[1]

Sensor-based remote monitoring could help better track Parkinson's disease (PD) progression, and measure patients' response to putative disease-modifying therapeutic interventions. To be useful, the remotely-collected measurements should be valid, reliable, and sensitive to change, and people with PD must engage with the technology. We developed a smartwatch-based active assessment that enables unsupervised measurement of motor signs of PD. Participants with early-stage PD ($N = 388$, 64% men, average age 63) wore a smartwatch for a median of 390 days. Participants performed unsupervised motor tasks both in-clinic (once) and remotely (twice weekly for one year). Dropout rate was 5.4%. Median wear-time was 21.1 h/day, and 59% of per-protocol remote assessments were completed. Analytical validation was established for in-clinic measurements, which showed moderate-to-strong correlations with consensus MDS-UPDRS Part III ratings for rest tremor ($\rho = 0.70$), bradykinesia ($\rho = -0.62$), and gait ($\rho = -0.46$). Test-retest reliability of remote measurements, aggregated monthly, was good-to-excellent (ICC = 0.75–0.96). Remote measurements were sensitive to the known effects of dopaminergic medication (on vs off Cohen's d = 0.19–0.54). Of note, in-clinic assessments often did not reflect the patients' typical status at home. This demonstrates the feasibility of smartwatch-based unsupervised active tests, and establishes the analytical validity of associated digital measurements. Weekly measurements provide a real-life distribution of disease severity, as it fluctuates longitudinally. Sensitivity to medication-induced change and improved reliability imply that these methods could help reduce sample sizes needed to demonstrate a response to therapeutic interventions or disease progression.

npj Digital Medicine (2022)5:65 ; https://doi.org/10.1038/s41746-022-00607-8



Polyp detection for colonoscopy



Examples of the variety of polyps detected by the ML system.

Image credit: Verily
Source: UEG Week

iGIE

Investigations

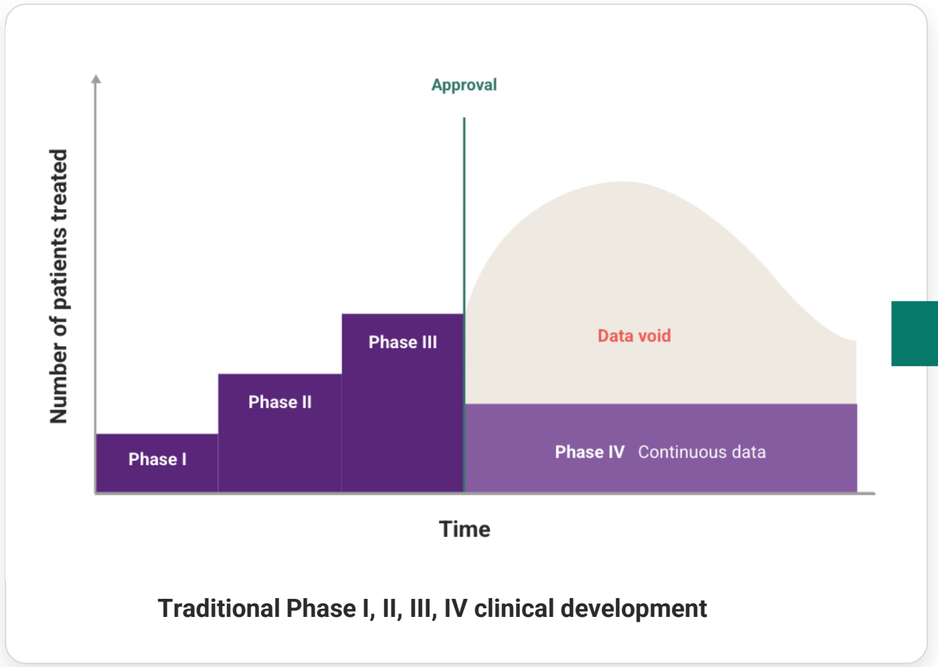
Novel artificial intelligence–enabled deep learning system to enhance adenoma detection: a prospective randomized controlled study

Jesse Lachter MD [1], Simon Christopher Schlachter PhD [2], Robert Scooter Plowman MD, MBA, MHSA, MSc [2 3], Roman Goldenberg PhD [2], Yaffa Raz MA [1], Nadav Rabani BSc [4], Natalie Aizenberg MSc [2], Alain Suissa MD [1], Ehud Rivlin PhD [2 5]

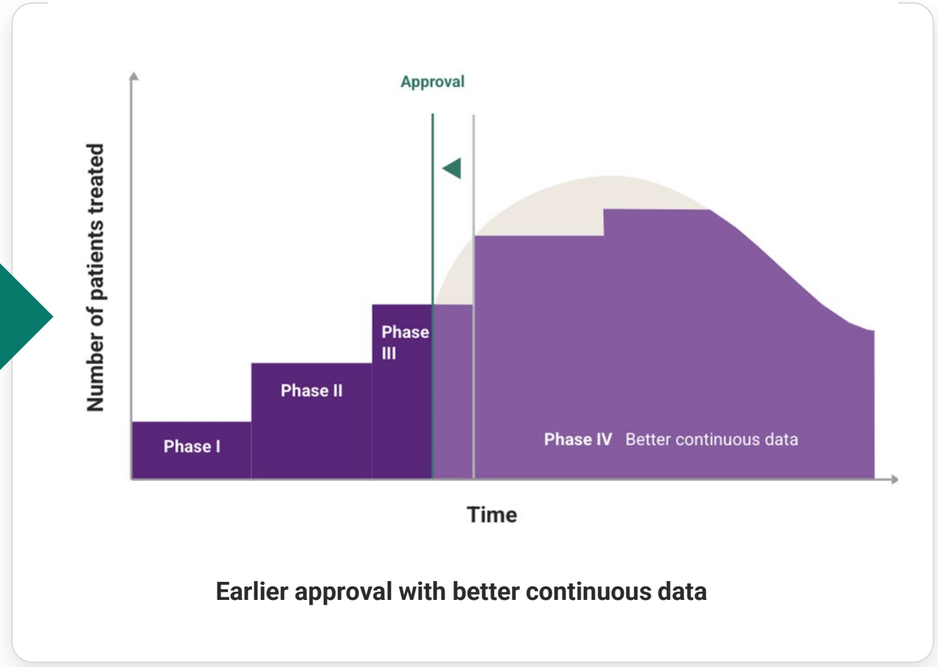
Background and Aims

Several artificial intelligence (AI) systems for polyp detection during colonoscopy have emerged in the gastroenterology literature and continue to demonstrate significant improvements in quality outcomes. This study assesses clinical quality outcomes during white-light colonoscopy with and without a novel AI computer-aided detection system, DEtection of Elusive Polyps (DEEP2), using Fuji 7000 series colonoscopes (Fujifilm, Singapore).

The shifting evidence generation landscape



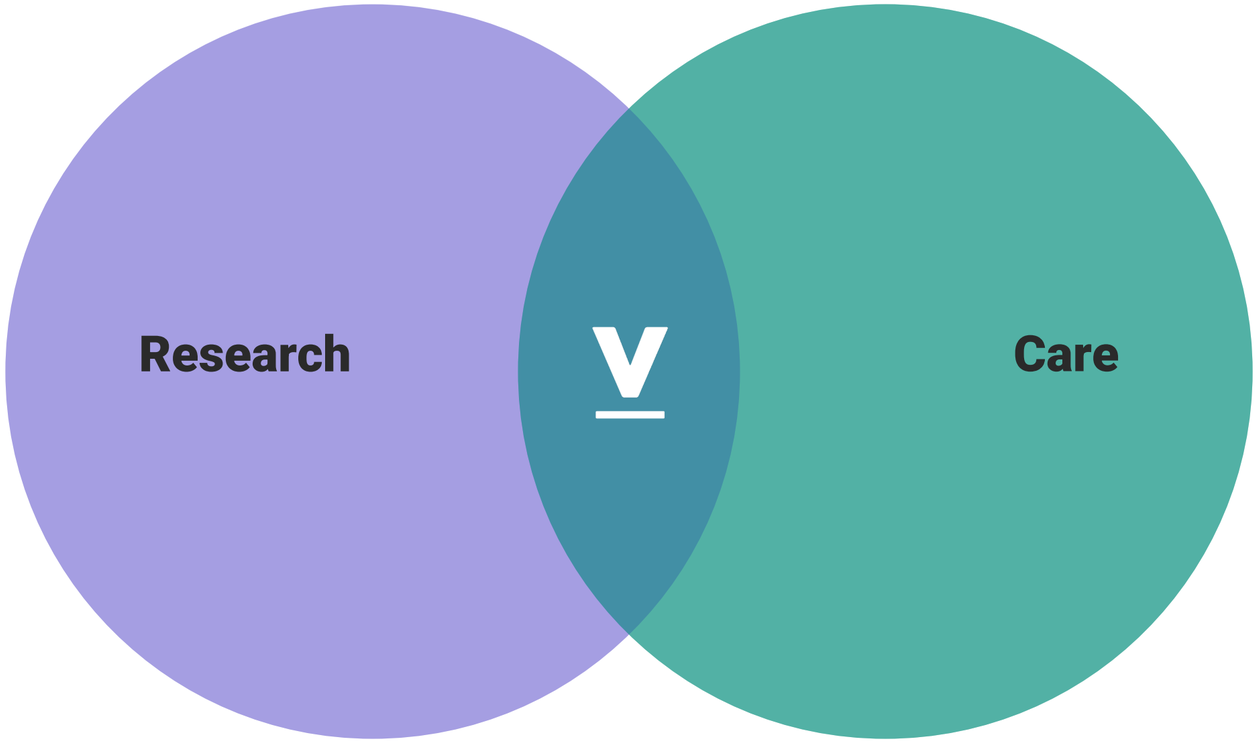
Traditional Phase I, II, III, IV clinical development



Earlier approval with better continuous data

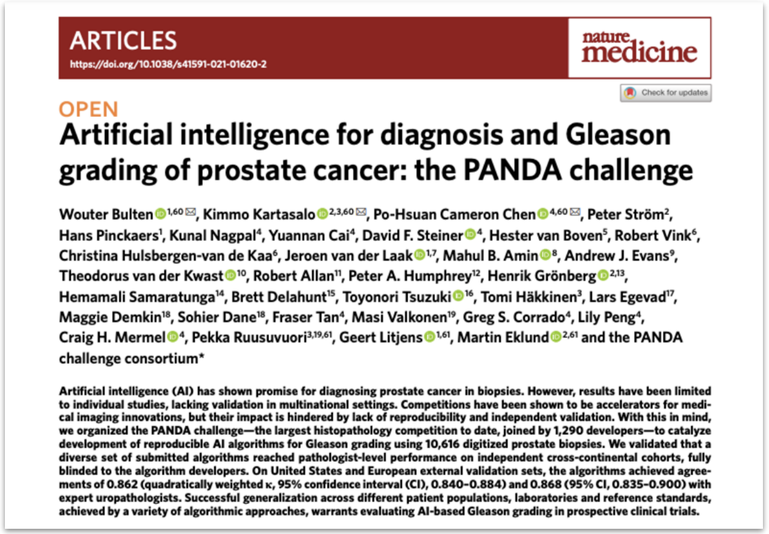
Building the evidence and data backbone that closes the gap between research and care

Accelerate the availability of safe and effective health solutions through faster, more patient-centric evidence generation



Help people live healthier lives through tools and insights that enable patients, providers & payers to make better decisions

Model development in oncology



ARTICLES
https://doi.org/10.1038/s41591-021-01620-2
nature medicine

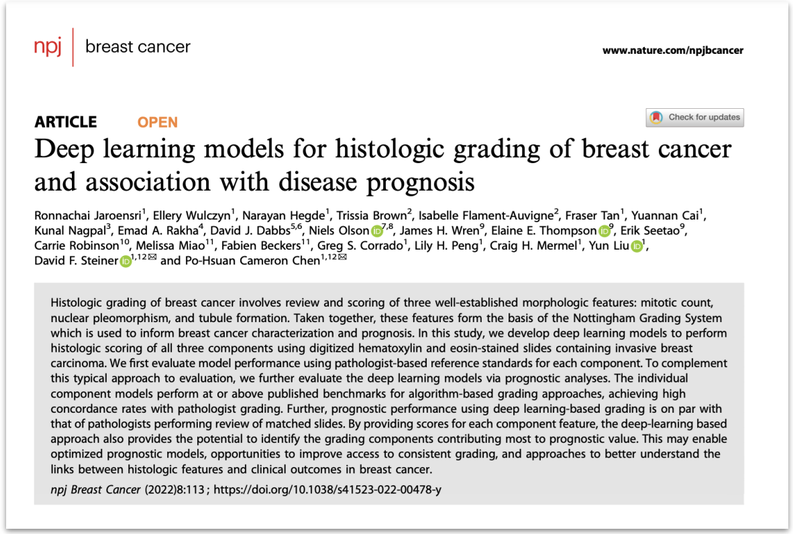
Check for updates

OPEN

Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge

Wouter Bulten[1,60], Kimmo Kartasalo[2,3,60], Po-Hsuan Cameron Chen[4,60], Peter Ström[2], Hans Pinckaers[1], Kunal Nagpal[4], Yuannan Cai[4], David F. Steiner[4], Hester van Boven[5], Robert Vink[6], Christina Hulsbergen-van de Kaa[6], Jeroen van der Laak[1,7], Mahul B. Amin[8], Andrew J. Evans[9], Theodorus van der Kwast[10], Robert Allan[11], Peter A. Humphrey[12], Henrik Grönberg[2,13], Hemamali Samaratunga[14], Brett Delahunt[15], Toyonori Tsuzuki[16], Tomi Häkkinen[3], Lars Egevad[17], Maggie Demkin[18], Sohier Dane[18], Fraser Tan[4], Masi Valkonen[19], Greg S. Corrado[4], Lily Peng[4], Craig H. Mermel[4], Pekka Ruusuvuori[3,19,61], Geert Litjens[1,61], Martin Eklund[2,61] and the PANDA challenge consortium*

Artificial intelligence (AI) has shown promise for diagnosing prostate cancer in biopsies. However, results have been limited to individual studies, lacking validation in multinational settings. Competitions have been shown to be accelerators for medical imaging innovations, but their impact is hindered by lack of reproducibility and independent validation. With this in mind, we organized the PANDA challenge—the largest histopathology competition to date, joined by 1,290 developers—to catalyze development of reproducible AI algorithms for Gleason grading using 10,616 digitized prostate biopsies. We validated that a diverse set of submitted algorithms reached pathologist-level performance on independent cross-continental cohorts, fully blinded to the algorithm developers. On United States and European external validation sets, the algorithms achieved agreements of 0.862 (quadratically weighted κ, 95% confidence interval (CI), 0.840–0.884) and 0.868 (95% CI, 0.835–0.900) with expert uropathologists. Successful generalization across different patient populations, laboratories and reference standards, achieved by a variety of algorithmic approaches, warrants evaluating AI-based Gleason grading in prospective clinical trials.



npj | breast cancer

www.nature.com/npjbcancer

ARTICLE OPEN

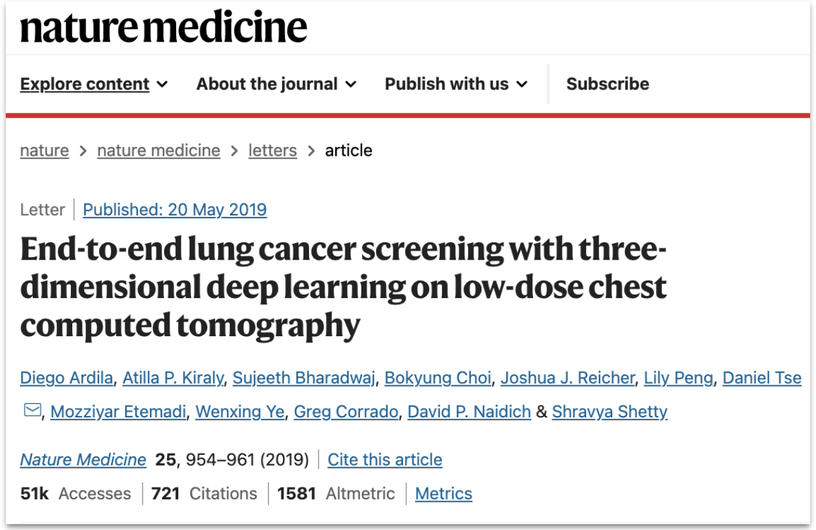
Check for updates

Deep learning models for histologic grading of breast cancer and association with disease prognosis

Ronnachai Jaroensri[1], Ellery Wulczyn[1], Narayan Hegde[1], Trissia Brown[2], Isabelle Flament-Auvigne[2], Fraser Tan[1], Yuannan Cai[1], Kunal Nagpal[3], Emad A. Rakha[4], David J. Dabbs[5,6], Niels Olson[7,8], James H. Wren[9], Elaine E. Thompson[9], Erik Seetao[9], Carrie Robinson[10], Melissa Miao[11], Fabien Beckers[11], Greg S. Corrado[1], Lily H. Peng[1], Craig H. Mermel[1], Yun Liu[1], David F. Steiner[1,12✉] and Po-Hsuan Cameron Chen[1,12✉]

Histologic grading of breast cancer involves review and scoring of three well-established morphologic features: mitotic count, nuclear pleomorphism, and tubule formation. Taken together, these features form the basis of the Nottingham Grading System which is used to inform breast cancer characterization and prognosis. In this study, we develop deep learning models to perform histologic scoring of all three components using digitized hematoxylin and eosin-stained slides containing invasive breast carcinoma. We first evaluate model performance using pathologist-based reference standards for each component. To complement this typical approach to evaluation, we further evaluate the deep learning models via prognostic analyses. The individual component models perform at or above published benchmarks for algorithm-based grading approaches, achieving high concordance rates with pathologist grading. Further, prognostic performance using deep learning-based grading is on par with that of pathologists performing review of matched slides. By providing scores for each component feature, the deep-learning based approach also provides the potential to identify the grading components contributing most to prognostic value. This may enable optimized prognostic models, opportunities to improve access to consistent grading, and approaches to better understand the links between histologic features and clinical outcomes in breast cancer.

npj Breast Cancer (2022)8:113 ; https://doi.org/10.1038/s41523-022-00478-y



nature medicine

Explore content | About the journal | Publish with us | Subscribe

nature > nature medicine > letters > article

Letter | Published: 20 May 2019

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

Diego Ardila, Atilla P. Kiraly, Sujeeth Bharadwaj, Bokyung Choi, Joshua J. Reicher, Lily Peng, Daniel Tse ✉, Mozziyar Etemadi, Wenxing Ye, Greg Corrado, David P. Naidich & Shravya Shetty

Nature Medicine **25**, 954–961 (2019) | Cite this article

Sources:

Performance Monitoring for AI/ML-based SaMD

5 Focus Areas:

1. Tailored Regulatory Framework for AI/ML-based SaMD
2. Good Machine Learning Practice (GMLP)
3. Patient-Centered Approach Incorporating Transparency to Users
4. Regulatory Science Methods Related to Algorithm Bias & Robustness
5. Real-World Performance (RWP)

Proactive monitoring of AI's real-world performance

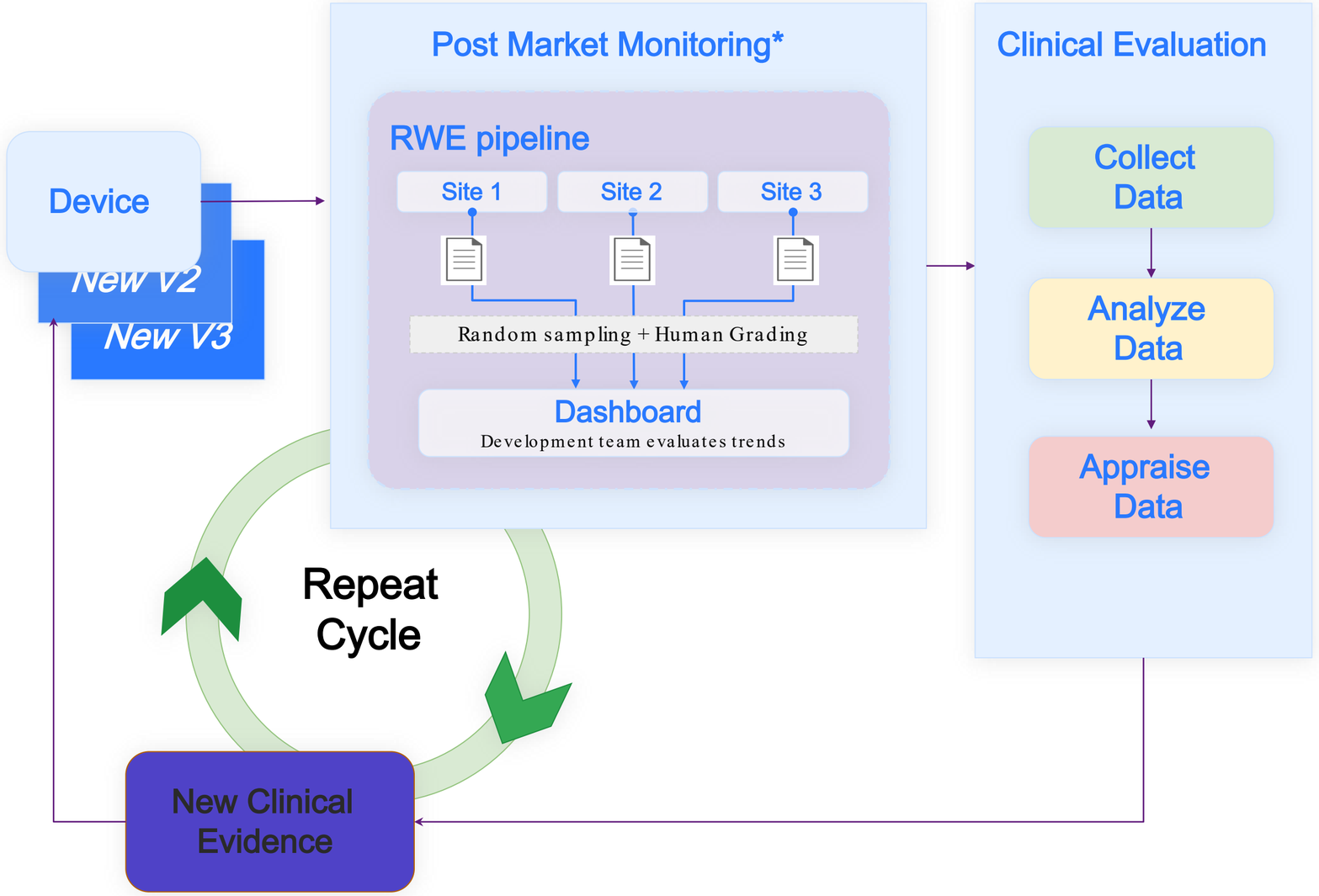
Pathway for Continuous Learning Leveraging Real World Performance Data

Intentional . Monitoring built in rather than an afterthought.

Embedded . Data collected during routine clinical care

Proactive vs reactive

Shared . Potential to aggregate across devices to develop a shared system to inform many stakeholders



Digital Measures in Cancer Care

Patient Reported Measures

Meaningful insights into physical, mental, and social health and quality of life (QoL). Increases HCWs awareness of symptoms.

Safety Monitoring

Early symptom recognition offers less complicated supportive care.

Treatment Performance

Response to treatment via longitudinal data provides meaningful clinical decision support.

Wearable Activity Metrics & Performance Status

Activity/Walking Analysis

Moderate to Vigorous Physical Activity (MVPA)

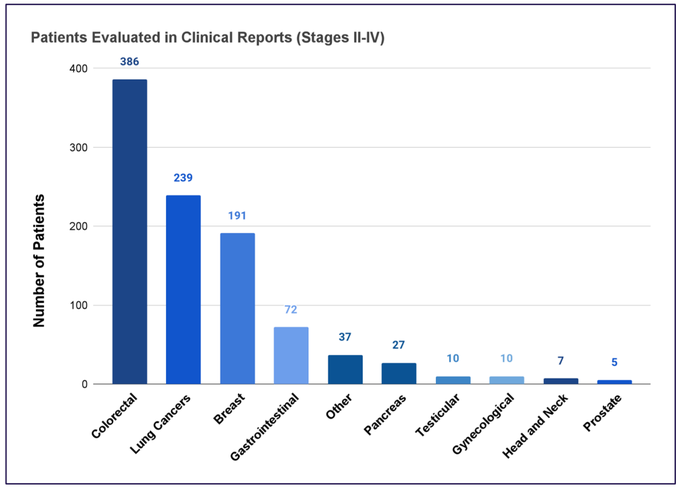
Activities of daily living (AODs)

Sleep Analysis

Performance Status

14

Clinical reports published that assessed Activity Metrics to inform Performance Status (2000-2021; *Kos et al, 2021*)



Statistically Significant Associations with ECOG Scores:

- Step Count
- Sedentary Behavior
- Posture
- Circadian rest/activity

verily

Thank you