Pathologist-led definition of generalisability and performance metrics for clinical AI ensures patient-relevant performance and clinical usability

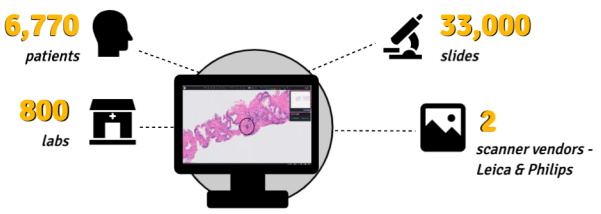
Patricia Raciti, MD; Peter Hamilton, PhD, FRCPath; Brandon Rothrock, PhD; Jillian Sue, MS; Margaret Horton, PhD; Christopher Kanan, PhD
1) Paige, 11 Times Square, 37th Floor, New York, New York, USA

Introduction

Artificial intelligence (AI) applied to diagnostic workflows promises accuracy and efficiency gains, and AI-based systems are available in the market for use today. "Generalisability" describes the ability for algorithms to perform across broad populations and on data from different laboratories, without refitting or calibration. Attaining and validating generalisability is not only a technical data problem but has clear clinical and patient safety implications. Furthermore, performance metrics must be clinically relevant and studied in subsets of disease states that are challenging for a human to interpret—and therefore, ripe for algorithmic assistance.

Materials and Methods

An AI-driven prostate cancer detection system was developed by Paige, using multiple-instance learning, that detects and indicates tissue suspicious for invasive cancer. The system was developed on diverse and clinically representative data from a single institution from >33,000 slides, >6770 patients.



In-silico **measures of success and generalisability do not necessarily translate into clinical performance, patient safety or pathologist usability metrics - whether the technology is safe and helpful for clinical end users or patients.**

Pathologist-Led Definition of Generalisability

- Safety
  - When algorithm is used by pathologists in independent clinical settings, diagnostic performance is improved.
  - When patient samples are analysed with algorithms across referral networks and systems, the algorithm performance is consistent.
- Workflow
  - The algorithm results presentation to diverse pathologists in independent settings positively impacts their decision-making and reporting.
  - Pathologists are consistently more efficient in their reporting and utilise additional tests optimally.

Data Science Led Definition of Generalisability

- Standalone *in-silico* performance criteria are met on cohorts of data external to training data
  - AUC
  - Sensitivity
  - Specificity

Future Work
- Health Economics studies on resource impact
- Prospective Studies
- Classifiers assessed on cancer mimics

Results

First, the system was validated on a validation set from >800 laboratories which included highly challenging cases. Additional validation confirmed performance across images acquired from different scanners. Then, the system was tested by independent pathologists at independent sites involving over 3,600 biopsies in different parts of the world. These pathologists were both generalists and GU specialised, and practiced both within the US and outside the US. On unseen data, without any site-specific calibration, the prostate cancer detection system performed with near-perfect sensitivity and high specificity, and was demonstrated to increase the diagnostic accuracy of pathologists.

Conclusion

While the development and initial in-silico validation was essential to creating the algorithm's technical ability to generalise, the performance on unseen prostate core biopsy data, from independent sites in different parts of the world, demonstrated the most relevant attributes of generalisability. A pathologist-led definition of generalisability and standardised criteria should be used to test new clinical AI applications in pathology. This will ensure that patient safety considerations are at the forefront of the clinical validation efforts for powerful but reliable AI applications.

Pathologist-led acceptance criteria based in clinically relevant performance metrics is crucial to understanding and improving pathologist+AI interaction, as well as for understanding the value of AI in the broader diagnostic workflow. Greater understanding of the strengths and weaknesses of a tool in the hands of a user, in addition to on a standalone basis, has the potential to increase trust, user satisfaction, and ultimately, patient outcomes in the form of correct, timely diagnoses.

References: Raciti P, Sue J, Ceballos R, Godrich R, Kunz JD, Kapur S, Reuter V, Grady L, Kanan C, Klimstra DS, Fuchs TJ. Novel artificial intelligence system increases the detection of prostate cancer in whole slide images of core needle biopsies. Mod Pathol. 2020 Oct;33(10):2058-2066. da Silva LM, Pereira EM, Salles PG, Godrich R, Ceballos R, Kunz JD, Casson A, Viret J, Chandarlapaty S, Ferreira CG, Ferrari B, Rothrock B, Raciti P, Reuter V, Dogdas B, DeMuth G, Sue J, Kanan C, Grady L, Fuchs TJ, Reis-Filho JS. Independent real-world application of a clinical-grade automated prostate cancer detection system. J Pathol. 2021 Apr 27. doi: 10.1002. Perincheri S, Levi AW, Celli R, Gershkovich P, Rimm D, Morrow JS, Rothrock B, Raciti P, Klimstra D, Sinard J. An independent assessment of an artificial intelligence system for prostate cancer detection shows strong diagnostic accuracy. Mod Pathol. 2021 Mar 29. doi: 10.1038/s41379-021-00794.