

DIGITAL PATHOLOGY & ARTIFICIAL INTELLIGENCE: FEASIBILITY FOR CLINICAL PRACTICE

Liron Pantanowitz

Original published work submitted to the Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Doctor of Philosophy (PhD) in Anatomical Pathology

Ann Arbor, Michigan, USA, 2022

Declaration

I, Liron Pantanowitz, student number 8701541J, declare that this Thesis is my own work and that I contributed adequately towards research findings published in the articles included in my Thesis. This Thesis is being submitted for the Degree of Doctor of Philosophy (PhD) in Anatomical Pathology at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other University.

Signature of Student:



31 August, 2022

.....
Liron Pantanowitz

.....
Date

Name of Primary Supervisor: Scott Hazelhurst

Signature of Supervisor:



31 August 2022

.....
Scott Hazelhurst

.....
Date

Name of Co-Supervisor: Pamela Michelow

Signature of Supervisor:



31 August 2022

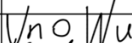
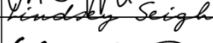
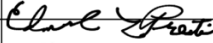
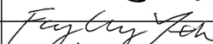
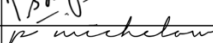
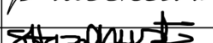
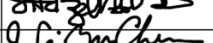
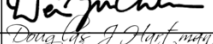
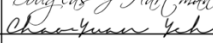
.....
Pamela Michelow

.....
Date

Agreement by co-authors: Given that this Thesis includes published articles with co-authors, the subsequent tables for each article include signatures of all co-authors

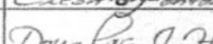
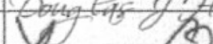
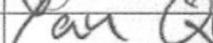
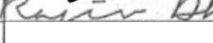
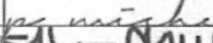
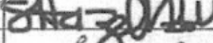
documenting their agreement for Liron Pantanowitz to use these published articles for his PhD degree. For all of the included published articles Liron Pantanowitz was the first and corresponding author, and primarily responsible for the study design, scientific work, manuscript preparation and successful submission for publication.

Article 1: Title: Artificial intelligence-based screening for mycobacteria in whole-slide images of tissue samples. *American Journal of Clinical Pathology*, 2021, volume 156(1), pages 117-128.

Authors	Name	Signature	Date
1 st author	Liron Pantanowitz		08/27/2020
2 nd author	Uno Wu		08/27/2020
3 rd author	Lindsey Seigh		08/28/2020
4 th author	Edmund LoPresti		08/27/2020
5 th author	Fang-Cheng Yeh		08/27/2020
6 th author	Payal Salgia		08/27/2020
7 th author	Pamela Michelow		08/31/2020
8 th author	Scott Hazelhurst		08/30/2020
9 th author	Wei-Yu Chen		08/31/2020
10 th author	Douglas Hartman		08/31/2020
11 th author	Chao-Yuan Yeh		08/28/2020

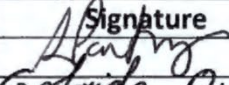
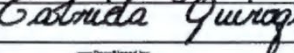
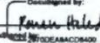
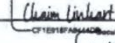
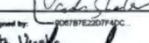
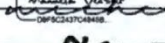
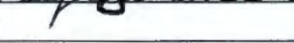
Note: 11 signatures of all co-authors are included in the above table.

Article 2: Title: Accuracy and efficiency of an artificial intelligence tool when counting breast mitoses. *Diagnostic Pathology*, 2020, volume 51(1), pages 80.

Authors	Name	Signature	Date
1 st author	Liron Pantanowitz		08/27/20
2 nd author	Douglas Hartman		08/28/20
3 rd author	Yan Qi		08/27/20
4 th author	Eun Yoon Cho		09/01/20
5 th author	Beomseok Suh		09/02/20
6 th author	Kyunghyun Paeng		09/02/20
7 th author	Rajiv Dhir		08/28/20
8 th author	Pamela Michelow		08/31/20
9 th author	Scott Hazelhurst		08/30/2020
10 th author	Sang Yong Song		09/01/2020
11 th author	Soo Youn Cho		09/01/2020

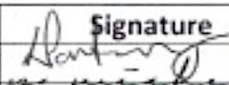
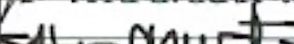
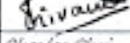
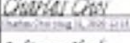
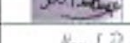
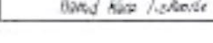
Note: 11 signatures of all co-authors are included in the above table.

Article 3: Title: An artificial intelligence algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies: a blinded clinical validation and deployment study. *The Lancet Digital Health*, 2020, volume 2, pages e407-416.

Authors	Name	Signature	Date
1 st author	Liron Pantanowitz		9/8/2020
2 nd author	Gabriela Quiroga-Garza		8/28/20
3 rd author	Lilach Bien		9/2/2020
4 th author	Ronen Heled		9/3/2020
5 th author	Daphna Laifenfeld		9/2/2020
6 th author	Chaim Linhart		9/3/2020
7 th author	Judith Sandbank		9/6/2020
8 th author	Anat Albrecht Shach		9/2/2020
9 th author	Varda Shalev		9/7/2020
10 th author	Manuela Vecsler		9/2/2020
11 th author	Pamela Michelow		08/31/20
12 th author	Scott Hazelhurst		08/30/2020
13 th author	Rajiv Dhir		08/28/20

Note: 13 signatures of all co-authors are included in the above table.

Article 4: Title: A digital pathology solution to resolve the tissue floater conundrum.
Archives Pathology & Laboratory Medicine, 2021, volume 145(3), pages 359-364.

Authors	Name	Signature	Date
1 st author	Liron Pantanowitz		8/31/2020
2 nd author	Pamela Michelow		08/28/20
3 rd author	Scott Hazelhurst		08/30/2020
4 th author	Shivam Kalra		Aug 31, 2020
5 th author	Charles Choi		Aug 31, 2020
6 th author	Sultaan Shah		Aug 31, 2020
7 th author	Morteza Babaie		Aug 31, 2020
8 th author	Hamid R. Tizhoosh		August 31, 2020

Note: 8 signatures of all co-authors are included in the above table.

Presentations arising from this study

1. **Validating AI apps for pathology practice**, Keynote address. 5th Digital Pathology & AI Congress. June 2019, New York City, NY, USA.
2. **Embedding AI into digital pathology workflows**. Digital Pathology Association (DPA) companion meeting. United States and Canadian Academy of Pathology (USCAP) annual meeting. March 2020, Los Angeles, CA, USA.
3. **Clinical validation and deployment of an AI-based algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies**. Wits Faculty of Health Sciences (FHS) research day & postgraduate expo. October 2020, virtual meeting.

Note: Awarded best student oral presentation in clinical sciences & therapeutics for health track.

4. **Artificial Intelligence and Anatomical Pathology: Feasibility for clinical practice**. Canadian Association of Pathologists. October 2020, webinar.
5. **AI – Is it ready for the clinic?** Swiss Society of Pathology 86th annual scientific meeting. November 2020, virtual plenary talk.
6. **Embedding AI into clinical practice: A pathologist's perspective**. European Society of Digital and Integrative Pathology (ESDIP). February 2021, virtual presentation.
7. **Artificial intelligence and the practice of Anatomical Pathology**. Papanicolaou Society of Cytopathology (PSC) companion meeting, March 2021, USCAP annual virtual meeting.
8. **Validating AI systems for clinical use in pathology practice**. AI-Med Clinician Series virtual conference. June 29, 2021.
9. **Validation of Artificial Intelligence-based tools in Digital and Computational Pathology**. Digital Pathology Association (DPA) webinar, September 30, 2021.
10. **Validating artificial intelligence for pathology practice**. TelemedEdu, USA. October 8, 2021. Troy, MI, USA.

Publications arising from this study

Article 1

Pantanowitz L, Wu U, Seigh L, LoPresti E, Yeh FC, Salgia P, Michelow P, Hazelhurst S, Chen WY, Hartman D, Yeh CY. Artificial intelligence-based screening for mycobacteria in whole-slide images of tissue samples. *American Journal of Clinical Pathology*, 2021; 156(1):117-128.
PMID: 33527136.

Article 2

Pantanowitz L, Hartman D, Qi Y, Cho EY, Suh B, Paeng K, Dhir R, Michelow P, Hazelhurst S, Song SY, Cho SY. Accuracy and efficiency of an artificial intelligence tool when counting breast mitoses. *Diagnostic Pathology*, 2020; 51(1):80.
PMID: 32622359

Article 3

Pantanowitz L, Quiroga-Garza GM, Bien L, Heled R, Laifenfeld D, Linhart C, Sandbank J, Albrecht Shach A, Shalev V, Vecsler M, Michelow P, Hazelhurst S, Dhir R. Clinical validation and deployment of an AI-based algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies. *Lancet Digital Health*, 2020; 2(8) e407-e416.
PMID: 33328045

Article 4

Pantanowitz L, Michelow P, Hazelhurst S, Kalra S, Choi C, Shah S, Babaie M, Tizhoosh HR. A digital pathology solution to resolve the tissue floater conundrum. *Archives of Pathology & Laboratory Medicine*, 2021; 145(3):359-364.
PMID: 32886759

Abstract

There is increasing interest in applying artificial intelligence (AI) tools to Anatomical Pathology. AI algorithms can detect rare events, automatically quantify features, and diagnose diseases by analyzing digital images. However, very few laboratories today routinely use such AI tools. Therefore, the aim of this study was to determine the feasibility of developing and validating AI technology for routine use in Anatomical Pathology. Four experiments were conducted that utilized whole-slide image datasets to train and test deep learning models. The first experiment involved a deep learning algorithm to screen digitized acid fast-stained slides for mycobacteria. With AI assistance pathologists were more accurate, quicker and found it easier to identify acid-fast bacilli than using manual modalities. The second experiment critiqued an AI tool that quantified mitotic figures in images of invasive breast carcinoma. For end-users of varying experience their accuracy and time spent counting mitoses improved with AI support. The third experiment concerned a blinded validation study and clinical deployment of an AI-based algorithm to aid reviewing prostate core needle biopsies. This algorithm was highly accurate at detecting prostate adenocarcinoma, distinguishing low- from high-grade Gleason scores, and identifying Gleason pattern 5 or perineural invasion. The fourth experiment demonstrated the success of using an AI-based image search tool to rapidly resolve the tissue floater conundrum encountered in pathology practice. All four AI-based tools successfully aided pathologists with routine tasks typically encountered in pathology practice and overall proved to be more accurate, efficient, standardized and easier to use than outdated and onerous manual methods.

Acknowledgements

Article 1

I thank Colleen Vrbín from Analytical Insights for her help with statistical analysis and Chitra Sharma and Cynthia Carbine from the University of Pittsburgh Medical Center in the USA for their administrative help. I thank the company aetherAI for their help in funding this study.

Article 2

I thank all of the participants in this study. I also thank Colleen Vrbín from Analytical Insights, LLC for her help with statistical analysis. I thank the company Lunit for their help in funding this study.

Article 3

I thank Ibex Medical Analytics for their help in funding this project.

Article 4

The results in this study are partly based on digital images offered for free by The Cancer Genome Atlas (TCGA) Research Network (<https://www.cancer.gov/tcga>). I thank the company Huron Digital Pathology for their financial support of this project.

Table of Contents

	Pages
Chapter 1: Introduction	1-34
Chapter 2: Published article 1	35-51
Chapter 3: Published article 2	52-64
Chapter 4: Published article 3	65-77
Chapter 5: Published article 4	78-86
Appendix 1: Ethics clearance certificate	87-88

List of Abbreviations

AI: Artificial intelligence
AFB: Acid-fast bacilli
AFS: Acid-fast stain
ASAP: Atypical small acinar proliferation
AUC: Area under ROC curve
C#: Computer programming language
CBIR: Content-based image retrieval
CI: Confidence interval
CNB: Core needle biopsy
CNN: Convolutional neural network
CPU: Central processing unit
FN: False negative
FP: False positive
GB: Gigabyte
GPU: Graphics processing unit
H&E: Hematoxylin and eosin stain
HER2: Receptor tyrosine-protein kinase erbB-2
HPF: High-power field
IBM: International Business Machines Corporation
IOU: Intersection Over Union
KB: Kilobyte
Ki67: Proliferation index antigen
mAP: Mean AP (area under precision recall curve)
Mdn: Median
MTb: *Mycobacterium tuberculosis*
NMS: Non-maximum suppression
NPV: Negative predictive value
OPT: Observer performance test
PCR: Polymerase chain reaction

PGY: Postgraduate year

PHH3: Phosphorylated Histone H3

PNI: Perineural invasion

PPV: Positive predictive value

PR: Precision recall

RAM: Random-access memory

RCNN: Regions with convolutional neural networks

SGD: stochastic gradient descent

ROC: Receiver operating characteristic

ROI: Regions of interest

SMC: Samsung Medical Center

TB: Tuberculosis

TCGA: The Cancer Genome Atlas

TP: True positive

TUPAC16: Tumor Proliferation Assessment Challenge in 2016

UPMC: University of Pittsburgh Medical Center

USA: United States of America

WSI: Whole slide image

ZN: Ziehl-Neelsen

List of Figures

Chapter (Article #)	Figure Number	Legend	Page
Chapter 2 (article #1)	1	Schematic showing acid-fast bacilli detection algorithm training workflow.	38
	2	Screenshot from the web portal showing regions of interest (patches) identified by the algorithm in the gallery on the left and the corresponding whole-slide image (WSI) on the right.	39
	3	Artificial intelligence–assisted detection of acid-fast bacilli (AFBs) is shown in a whole-slide image (WSI).	41
	4	Examples of artificial intelligence–detected acid-fast bacilli (AFB) structures.	42
	5	Area under the curve (AUC) for acid-fast bacilli algorithm detection in image patches.	42
	6	Proportion of slides positive for acid-fast bacilli (AFBs) screened by different modalities.	43
	7	Review time by screening modality.	44
	8	Pathologist-perceived difficulty of identifying acid- fast bacilli by screening modality.	45
Chapter 3 (article #2)	1	Flow chart of the methodology and datasets employed in developing and validating an AI-based tool to quantify mitoses in breast carcinoma.	55
	2	Web-based tool showing a HPF of breast carcinoma.	56
	3	Algorithm performance for mitotic figure detection in the analytical validation dataset.	57
	4	Accuracy and precision with and without AI support per user experience level.	58
	5	Median number of seconds spent with and without AI support per user experience level.	59
Chapter 4 (article #3)	1	Overview of the algorithm and clinical deployment of the Galen Prostate second read system.	68
	2	Examples of diagnoses after review.	70
	3	Missed cancer case originally diagnosed as benign.	71

Chapter (Article #)	Figure Number	Legend	Page
Chapter 5 (article #4)	1	Fabricated slide containing a section of renal cell carcinoma and 2 adjacent separate colon cancer and bladder cancer tissue floaters.	80
	2	Schematic illustration of the general idea of using barcodes for image representation: whole slide image indexed by converting separate patches into barcodes.	81
	3	Schematic diagram showing how the origin of a suspected floater gets detected.	82
	4	Indexing of a sample whole slide imaging (scan with bladder tumor) (B) yielding 33 patches to build a mosaic.	83

List of Tables

Chapter (Article #)	Table Number	Title	Page
Chapter 2 (article #1)	1	Breakdown of Whole-Slide Images (WSIs) Used in Acid-Fast Bacilli Algorithm Development	37
	2	Algorithm Performance at Image Patch and WSI Levels	42
	3	Comparison of Algorithm Performance Using Individual or Combined Convolutional Neural Network Models	42
	4	Sensitivity, Specificity, PPV, PPV, and Accuracy by Reviewer and Method	43
	5	Comparison of Time Category for Pathologists by Review Method (n = 138)	44
	6	Summary of Published Studies Using Image Analysis to Identify Mycobacteria	46
Chapter 3 (article #2)	1	Profile of invasive ductal carcinoma cases enrolled in the study	55
	2	Accuracy by experience level	58
	3	True positive (TP), false positive (FP), and false negative (FN) values for mitotic cell detection	59
	4	Median time to count mitoses by study participant experience level	60
Chapter 4 (article #3)	1	Algorithm performance	68
	2	Pathologists' misdiagnoses identified by the algorithm	69
	3	Performance of algorithms in detection and grading of prostate cancer	71
Chapter 5 (article #4)	1	Top 20 Primary Sites With the Highest Number of Whole Slide Imaging (WSI) in the Dataset	80
	2	Initial Matched Results for Each Tissue Floater (UPMC 300 Whole Slide Imaging Dataset)	81
	3	Image Search Results That Match Tissue Floaters (UPMC p NCI 2325 Whole Slide Imaging Dataset)	82
	4	Top 5 Retrieved Patches for Manual Search	83

CHAPTER 1: INTRODUCTION

1.1 DIGITAL PATHOLOGY

The application of Artificial Intelligence (AI) in Anatomical Pathology requires digital images and a suitable Information Technology (IT) infrastructure. For this reason, a brief overview of Digital Pathology is necessary. Digital imaging has benefited the field of Pathology for clinical practice, research and education [1]. Digital data is often easier to share, archive, integrate and analyze. Digital images also enable pathologists to employ next generation tools such as image analysis [2]. Today, many pathology laboratories are using whole slide imaging (WSI) [3]. WSI is the process whereby a pathology glass slide with tissue from a patient can be digitized (scanned) using a whole slide scanner and the acquired whole-slide image (digital slide) can be viewed on a computer monitor [4]. Numerous commercial vendors offer hardware and/or software solutions for WSI. There are multiple clinical (e.g. telepathology) and non-clinical (e.g. research, education) applications. Furthermore, in several countries there is regulatory approval of this technology for clinical diagnostic use [5]. Nevertheless, whilst there is increasing interest among pathologists to use WSI the global adoption of Digital Pathology has been slow due to technological, financial, regulatory and cultural barriers [6]. Recent efforts to couple WSI with advanced computation and deep learning methods have incited much interest in applying AI to pathology, which in turn is driving the field of Digital Pathology forward.

1.2 ARTIFICIAL INTELLIGENCE IN PATHOLOGY

AI is a branch of computer science that is currently very topical. Recently, there has been much hype regarding the use of AI tools applied to healthcare [7]. Over the last two decades digital imaging technology and its application in pathology have increased greatly, which has ushered in a new field called Computational Pathology [8]. WSI has enabled the generation of large digital datasets in the Pathology community that, in turn, facilitates the development and validation of AI algorithms. AI has been applied to pathology for both clinical purposes and discovery. AI-based tools have demonstrated outstanding performance, but this has been mostly with narrow (weak) AI systems focused on limited tasks (e.g. mitotic figure detection). A plethora of studies about AI using deep learning in pathology have been published showing novel applications [9-14]. These algorithms have been used to detect rare events, automatically quantify features in digital images, diagnose diseases (e.g. cancer) from WSI, and even make prognostic predictions by analyzing pixels.

Clearly, there is great potential for leveraging digital pathology to support AI [15]. AI in pathology offers better diagnostic accuracy and more precise measurements compared to those rendered by humans, promotes standardization, and also permits automation. AI can accordingly benefit pathologists practicing in most settings, including low middle income countries. For example, there is a dearth of anatomical pathologists in South Africa [16]. If pathologists in South Africa could exploit AI to handle more routine and

mundane tasks this could free them up to perform more important work and allow them to rather focus their attention on more difficult cases.

It has been pointed out in the literature that most studies evaluating AI applications in healthcare to date have not been vigorously validated for reproducibility, generalizability and safety in the clinical setting [17-19]. (Bi et al 2019; Maddox et al 2019). Despite advances with AI technology applied to pathology there are currently very few laboratories routinely using these tools. Studies demonstrating the feasibility of using AI to assist pathologists in daily practice are thus limited. Addressing this gap in the field would make a substantial contribution to advancing this field and provide much needed translational evidence that could help develop recommendations and guidelines for the safe and effective use of AI in routine diagnostic Anatomical Pathology workflow. The experiments discussed herein demonstrate the feasibility of employing AI-based tools to successfully solve various clinical challenges in Anatomical Pathology.

1.3 STUDY GOALS

The aim of this study was to determine the feasibility of implementing AI technology for routine use in Anatomical Pathology. Considerable literature has accumulated showing the potential for AI to assist pathologists. However, it is unclear to what extent AI can support pathologists with everyday practical tasks. AI-based tools for four different clinical scenarios in Anatomical Pathology were consequently deployed in a modern clinical laboratory setting with adequate computational capacity and pathology

informatics expertise to demonstrate that AI can assist pathologists in performing specified tasks.

Four challenges in pathology, that are often time-consuming to undertake manually, and that were intended to be solved by AI-based methods included:

- **Experiment #1:** Screening digital slides for mycobacteria
- **Experiment #2:** Grading breast carcinoma using mitotic figure counts
- **Experiment #3:** Screening prostate core needle biopsies (CNBs) for cancer
- **Experiment #4:** Resolving the original source of contaminating tissue floaters sometimes encountered on pathology slide

In the first experiment, the objective was to develop and validate a novel AI tool to screen digitized acid-fast stained (AFS) slides in order to detect mycobacterial infection.

In the remaining experiments, the objectives were to clinically validate previously developed algorithms on external datasets to either quantify mitotic figures in breast carcinoma, diagnose and grade prostate adenocarcinoma, or utilize a search tool to query for matching regions of interest (ROI) in whole-slide images.

1.4 MATERIALS AND METHODS SYNOPSIS

The four experiments in this study were primarily conducted at the University of Pittsburgh Medical Center (UPMC) in Pittsburgh, USA, where Liron Pantanowitz was employed. Ethics clearance was accordingly obtained from both the Human Research

Ethics Committee (Medical) of the University of the Witwatersrand (Appendix 2) and the Institutional Review Board (IRB) of the University of Pittsburgh, USA.

All experiments utilized whole-slide images acquired by scanning archival glass slides, typically at 40x magnification (0.25 μm /pixel resolution) with a single Z-focus plane. A variety of commercial whole slide scanners were used including Aperio (Leica Biosystems, USA), Philips (Philips, Netherlands), Nanozoomer (Hamamatsu, Japan) or 3D Histech (3DHISTECH, Hungary) instruments. For experiment 4, an additional digital dataset was downloaded from The Cancer Genome Atlas (TCGA) public archive. Digital datasets were de-identified and when needed coupled with available clinical metadata such as patient demographics, pathology diagnoses, and ancillary laboratory test results.

Development of deep learning models followed the process delineated by Harrison and colleagues that included data acquisition, expert annotation (labeling), data partitioning into training and test data, algorithm training, blinded validation on at least one independent dataset, and verification prior to deployment [13]. Whole-slide images (typically 100,000 x 100,000 pixels) were broken into smaller patches or tiles (e.g. 64 x 64 pixels) with no overlap and subsequently fed into one or multiple multilayered convolutional neural networks (CNN). For experiment 4, an ensemble approach utilized a cohort of different algorithms that exploited both supervised and unsupervised computational methods for image processing. For certain experiments, data augmentation techniques (e.g. random rotation, jittering to add small amounts of noise)

were applied. Different hardware (e.g. CPUs and GPUs), software solutions (e.g. Ubuntu) and programming languages (e.g. Python, C/C++, JavaScript) were utilized for each experiment. Algorithm output at the image patch level, and occasionally at the WSI level, was conveyed as explainable annotations, heatmaps overlaid on hematoxylin and eosin (H&E) digital images, or as image patches ranked and displayed in gallery format.

The results of different AI tools were compared to manual microscopic interpretations rendered by humans. The ground truth for resolving discrepancies was established by expert pathologist reading, consensus review and/or the results of ancillary tests (e.g. immunohistochemical staining, microbiology culture results, molecular findings). Metrics used to evaluate AI model performance included accuracy, sensitivity/specificity, positive/negative predictive value, F-score, area under the receiver operating characteristic curve (AUC), precision recall curve, and/or recorded time to perform a specific task.

1.5 SYNTHESIS OF EXPERIMENTS

1.5.1 EXPERIMENT #1: Acid-Fast Bacilli Screening AI Tool

A detailed report of this experiment is provided in Chapter 2 [18].

1.5.1.1 Background

Mycobacteria cause significant disease worldwide. In South Africa, for example, annually hundreds of thousands of people fall ill with tuberculosis and thousands die from this infectious disease [19]. When anatomical pathologists encounter pathologic changes in tissue specimens (e.g. granulomatous inflammation) that may be caused by mycobacterial infection they are obliged to try identify these microorganisms concealed within these samples. This typically requires the use of ancillary studies. A simple and relatively cheap method is to stain tissue sections with an AFS (e.g. Ziehl-Neelsen or Kinyoun stain) to highlight acid-fast bacilli (AFB) such as mycobacteria. Since AFB are small (e.g. only 2-4 μm in length) and may be sparse, it is cumbersome for a busy pathologist to screen stained slides. Therefore, the possibility of computer-assisted screening for this mundane task is attractive.

To date, publications about image algorithms to detect AFB have been restricted mostly to sputum samples and exploited static snapshots [18, 20]. To the best of my knowledge, very few studies have reported developing a deep learning model to analyze entire digital slides in order to identify mycobacteria in human tissue [21, 22]. These prior studies used small datasets for algorithm training. The objective of this first experiment was to thus develop and validate a comparable deep learning algorithm, but with a much larger dataset, to screen digitized AFS slides for mycobacteria within tissue sections, and to subsequently deploy this AI-based tool for use in clinical practice.

1.5.1.2 Materials and Methods

A deep learning algorithm was developed using a sizeable digital dataset (n=441) comprised of scanned positive and negative AFS slides from UPMC (USA) and Wan Fang Hospital (Taiwan). Numerous ROIs containing either AFB (n = 6,817), mimics (n = 7,426), or negative background areas (n= 3,601) were annotated in these whole-slide images by several pathologists. Digital slides were subsequently cropped into thousands of image patches to train two CNN models. The algorithm's output was presented in the form of image thumbnails ranked according to the likelihood that they contained at least one possible AFB. The thumbnails were displayed online in gallery format, together with their corresponding digital slide, for a pathologist end-user to review. After the algorithm was successfully validated, the accuracy and speed (measured in seconds) of two pathologists at finding AFB in an independent dataset (n=138) were compared using manual light microscopy, whole-slide image evaluation without AI support, and web-based AI-assisted analysis.

1.5.1.3 Results

The algorithm demonstrated excellent performance (AUC = 0.960 at the image patch level; AUC = 0.900 at the digital slide level) when compared to the ground truth (i.e. original "signed-out" diagnosis rendered by a pathologist). The AI-based method was found to be more sensitive and accurate than humans alone. Pathologists were accordingly also able to identify significantly more AFB with AI-assistance than manual

microscopy or WSI examination without the aid of AI. Moreover, screening for mycobacteria proved to be significantly quicker and easier with AI-assistance.

1.5.1.4 Conclusion

This experiment successfully demonstrated the feasibility of using an AI-based tool to search for rare events, such as AFS mycobacteria in whole-slide images. Compared to alternate manual methods of screening, with AI assistance this onerous task was more sensitive, accurate, and quicker to perform.

1.5.2 EXPERIMENT #2: Breast Mitosis Quantification AI tool

A comprehensive report of this experiment is provided in Chapter 3 [23].

1.5.2.1 Background

The proliferation activity in breast carcinoma, which relies on mitotic count, is an important prognostic marker. Breast cancer grading accordingly requires that a pathologist count mitotic figures in H&E stained histology sections of a patient's breast cancer sample. However, manually counting mitotic figures is subjective, time-consuming, and suffers from low reproducibility. Several articles were published showing that digital image analysis can solve this challenge by automating mitosis quantification [24-25]. However, early efforts leveraging computer-assisted mitotic

counting were limited to using only pre-defined ROIs in images. More recent work has resorted to developing automated methods using deep learning techniques to predict tumor proliferation from whole-slide images [26-27].

To date, there have been no studies showing whether using an AI tool to detect and quantify mitoses in breast carcinoma actually improves end-user accuracy and efficiency when scoring mitotic figures in practice. The aim of this study was to accordingly critique such an AI tool for use in clinical practice. The hypothesis was that reviewer accuracy and efficiency would improve with AI support.

1.5.2.2 *Materials and Methods*

All components of this study were conducted at two different medical centers, UPMC in the USA and Samsung Medical Center (SMC) in Seoul, South Korea. This helped increase the dataset size and heterogeneity for algorithm training, as well as limit bias in the reader study used to assess the AI system's practicability. Representative H&E glass slides from breast cancer cases (n = 320) were scanned at 40x magnification at UPMC using an Aperio AT2 scanner (Leica Biosystems, USA) and at SMC using a 3D Histech P250 scanner (3DHISTECH, Hungary). To simplify working with large image files, the whole-slide images were broken up into representative digital patches called high-power fields (HPFs). To further train the algorithm developed by the company Lunit, 10 expert pathologists were recruited to label mitotic figures. To assess the efficacy of this algorithm a reader study involving 24 end-users of varying proficiency

was undertaken. After receiving informed consent, the readers were asked to count mitotic figures in 140 HPFs with and without the aid of AI. Their accuracy and time (measured in seconds) to perform this task were recorded using a web-based tool.

1.5.2.3 Results

The accuracy, precision and sensitivity of counting mitotic figures in images of invasive breast carcinoma for pathology end-users improved when they were provided with AI support to perform this task. With AI support, most readers (87.5%) were also able to detect more mitoses and had fewer falsely flagged mitotic figures. Of note, with AI assistance there was better standardization among readers, manifested by higher inter-pathologist agreement. Moreover, AI assistance reduced the overall time (27.8% savings) that end-users spent on this task.

1.5.2.4 Conclusion

This study validated that not only can an AI algorithm successfully perform a narrow task such as counting mitotic figures in digital images of invasive breast carcinoma, but that this AI-based tool can augment pathology end-users when performing this mundane task. End-users with AI assistance were more accurate, efficient and less varied at quantifying mitotic figures.

1.5.3 EXPERIMENT #3: Prostate Cancer Diagnosis AI Tool

A complete report of this experiment is provided in Chapter 4 [28].

1.5.3.1 *Background*

In Anatomical Pathology practice, screening prostate CNBs for carcinoma is challenging. In many practices these patient samples are typically received in high volume, screening is time-consuming, foci of cancer may be very small, there is notable inter-observer variability especially with grading, and there are several mimics that can lead to a misdiagnosis. To help address this challenge, several investigators have developed deep learning algorithms capable of analyzing digital pathology images of prostate histopathology [29-32].

Most such algorithms published in the literature have been limited to performing narrow tasks (e.g. only cancer detection or grading). However, pathologists are required to perform multiple tasks when evaluating a prostate CNB such as identify, grade and quantify prostate adenocarcinoma. Pathologists also have to report about additional findings important for patient management such as the presence of any Gleason pattern 5 which is a strong predictor of poor outcome and perineural invasion (PNI) which is a predictor of extra-prostatic tumor extension. To date, very few AI tools capable of comprehensively analyzing prostate CNBs have been deployed and validated for use in routine clinical practice. The aim of this study was to address this gap.

1.5.3.2 Materials and Methods

A deep learning algorithm was developed to diagnose prostate adenocarcinoma, grade and quantify detected cancer, as well as find PNI in H&E stained slides of prostate CNBs that were scanned using a Philips scanner (Philips, Netherlands). Labeled whole-slide images used for this purpose included a training dataset of 549 slides further broken into 1,357,480 image patches, and a separately held-out test dataset of 2,501 digital slides. Following calibration, the algorithm was subsequently validated on an external dataset of 100 consecutive archival cases comprised of 1,627 scanned slides using an Aperio AT2 scanner (Leica Biosystems, USA). Discrepancies between algorithm and original pathologist's diagnoses were managed by blinded review and the consensus of three expert pathologists, as well as immunohistochemistry (PIN4 cocktail stain that consisted of p63 + CK-903 + P504S) to confirm the diagnosis. This AI-based system was implemented in the clinical pathology laboratory at Maccabi Healthcare Services in Israel, and deployed to retrospectively review all prostate CNBs as a quality control application (i.e. second read system).

1.5.3.3 Results

The algorithm performed exceptionally well on both internal and external datasets. For identifying prostate adenocarcinoma the AUC was 0.991 in the external validation set. For distinguishing between Gleason low- and high-grade cancers the AUC was 0.941 in

the external validation set. For detecting Gleason pattern 5 the AUC was 0·971 in the external validation set. For identifying PNI the algorithm achieved an AUC of 0·957 in the external validation set. As a second read system in the Maccabi pathology laboratory, this AI-based tool provided numerous helpful alerts for potential disagreements and correctly identified one case of missed adenocarcinoma among 941 reviewed cases.

1.5.3.4 Conclusion

This study reports, to the best of my knowledge, the first successful development, external clinical validation, and deployment in clinical practice of an AI-based algorithm to accurately detect, grade, quantify and assess for PNI in whole-slide images of prostate CNBs.

1.5.4 EXPERIMENT #4: Pathology Image Search AI Tool

A detailed report of this experiment is provided in Chapter 5 [33].

1.5.4.1 Background

In routine clinical practice pathologists sometimes encounter extraneous pieces of tissue on slides that may be due to specimen cross contamination. These are often called “tissue floaters”. Figuring out if they belong to the patient’s case being reviewed

or whether this tissue fragment arose from another case is cumbersome and time-consuming, even if expensive molecular techniques are used [34-35]. The aim of this study was to develop an AI-based image search tool to resolve the tissue floater conundrum.

1.5.4.2 *Materials and Methods*

A glass slide with tissue floaters containing different cancers was artificially created by adding two separate H&E stained tissue fragments on the edge. This constructed slide was scanned using an Aperio AT2 scanner (Leica Biosystems, USA), along with the two original cancer slides used to create these tissue floaters. These digital slides were then randomly embedded into a large whole-slide image dataset ($n = 2,325$), encompassing many different H&E stained pathology entities. UPMC provided 300 images, whereas most of these images ($n = 2,025$) were obtained from the public dataset offered by The Cancer Genome Atlas (TCGA). Whole-slide images were tagged with their pathology diagnosis and the affected anatomic site. Image patches derived from these whole-slide images were converted into linear barcodes and indexed for easy retrieval. A deep learning-based image search tool using an ensemble approach of different algorithms was then utilized to mine for matching image features via barcodes. The retrieved images, each ranked based on their likelihood of matching the queried image, were sorted and then displayed in gallery format. The accuracy of matching the original tumor images using this model to each tissue floater was recorded.

1.5.4.3 Results

The image search tool performed as intended and required only milliseconds to complete a query. When the digital database was repeatedly queried using this tool the likelihood of retrieving the correct tumor match to the tissue floater was very high.

Further, the accuracy of retrieving the correct matching image increased when successively greater regions of the tissue floater were chosen.

1.5.4.4 Conclusion

The AI-based model developed was successful for content-based image retrieval (CBIR). This image search tool offers pathology laboratories, especially those that have transitioned to going fully digital with WSI, a novel method to rapidly resolve the tissue floater conundrum.

1.6 DISCUSSION

AI encompasses a variety of techniques devised to mimic human intelligence. The application of AI in pathology is relatively new [13, 36-38]. Nevertheless, this innovative field is rapidly evolving due to advances in digital pathology (e.g. WSI), availability of large curated digital datasets, better computation (e.g. graphics processing units, cloud computing), and progress in machine learning (e.g. deep learning networks that contain multiple hidden layers of nodes). Deep neural networks can now be trained to identify

precise objects, larger ROIs, or even analyze entire whole-slide images. Of the diverse artificial neural networks available, CNNs have been most commonly applied to developing pathology image classification systems [13, 39]. Several global machine learning competitions (e.g. CAMELYON and TUPAC challenges) that exploited CNNs helped promote the role of AI in pathology [40-41]. Whilst most early work in this field was largely confined to research laboratories, recently several startup AI companies (e.g. aetherAI, Lunit, Ibex, PathAI, Paige) have emerged offering the pathology community promising commercial AI-based tools intended for clinical and/or research use. According to a survey published in 2019 involving 487 pathologists practicing in 54 countries, the vast majority (80%) of respondents expected AI tools to soon be incorporated into pathology laboratories [42]. There are many benefits to adopting AI in pathology. These include eliminating tedious tasks, improved accuracy, standardization, predicting outcomes, better efficiency, reduced workload, and automation [43]. Hence, there is much enthusiasm about implementing AI in routine clinical practice.

The study herein reports four separate experiments that successively demonstrate the capabilities of AI in Anatomical Pathology. Each experiment culminated in a peer-reviewed publication that made important contributions to the field. **Article #1** regarding an AI tool used to screen for AFB in digital slides that was published in the journal *Archives of Pathology & Laboratory Medicine* (impact factor 4.094) was accessed 2,534 times online, had 784 PDF downloads, and was shared 18 times on social media only four months after being published. **Article #2** concerning the mitotic figure quantification AI tool for breast carcinoma that was published in *Diagnostic Pathology* (impact factor

2.192) was accessed 2,582 times online 14 months since its publication and cited 8 times according to Web of Science. According to Google Scholar, **article #3** about the AI tool used to analyze prostate cancer has been cited 47 times in the one year since it was published in *The Lancet Digital Health* (impact factor 24.519). There was also an editorial published alongside this paper in which the authors stated “The resulting study is distinguished by its real-world evaluation, showing how computer-aided diagnosis (CAD) tools might influence pathology practice in the near future”[44]. **Article #4** that was published in the journal *Archives of Pathology & Laboratory Medicine* (impact factor 4.094) was accessed 2,543 times online, had 785 PDF downloads, and was tweeted on Twitter by 30 individuals in the 14 months since being published.

Experiments 1, 2 and 3 replicated specialized tasks pathologists manually perform with their microscope, whereas experiment 4 achieved a task pathologists are currently typically unable to accomplish alone using a traditional light microscope. Experiments 1 and 2 demonstrated the ability of an AI-based tool to perform a narrow (specific) task (i.e. screen for rare mycobacteria or detect and quantify mitotic figures in images). Indeed, computers are better suited than humans to screen large image files for rare events, such as a single mycobacterium that only occupies a minute portion (e.g. 10 x 10 resolution) of an entire whole-slide image (e.g. 100,000 x 100,000 resolution) [18]. Experiment 3 successfully demonstrated the ability of an AI-based tool to simultaneously perform multiple tasks (i.e. identify, grade and characterize additional features about prostate cancer). Experiment 4 established the ability of an AI-based search tool to rapidly query a large image dataset for matching pathology images.

While many AI tools in pathology have been theoretically shown to be valuable, most of them have not yet been deployed widely in clinical practice. They have accordingly not been validated in real-world clinical settings for reproducibility and generalizability. Therefore, best practices for implementation (e.g. validation studies, underlying IT requirements, pathologist oversight needed) and the impact of AI tools on end-users remains to be defined. Each of the aforementioned published experiments offers some evidence to address this gap, which could be used in the future development of standard clinical guidelines. Experiment 1, for example, provided information about the optimal washout period to be used during implementation when comparing AI versus manual modalities. While most published data insinuates that pathologists working with AI tools typically outperform humans or machines separately [13], their optimal setup (e.g. workflow, end-user interaction) in practice still needs to be ascertained. For example, experiment 2 aptly highlighted that not all end-users may interact with AI-tools as anticipated, which could be addressed with more training and ongoing performance assessment. Experiment 3, on the other hand, highlighted the importance of calibrating algorithms locally pre-deployment.

All of the experiments in this study followed a similar process of developing AI algorithms that included training followed by validation using an independent internal and/or external dataset. For experiments 2 and 3, roughly half of the data collected was used for training and the rest held-out for testing. To avoid bias, it was important to ensure that both of these datasets had a similar distribution of cases and features. The

strategy employed for training in experiments 1, 2 and 3 was supervised learning, where delineated data inputs were used to predict matching outputs. This involved pathology experts assigning labels (e.g. pathology diagnoses) to images and/or manually annotating specific features (e.g. mycobacteria, mitotic figures, prostate glands, etc.) within images. While manual training was effective, it was labor-intensive, time-consuming and sometimes subject to observer variability. Recently, some investigators have advocated using weak supervision in which they combine small datasets that have detailed annotations with larger unlabeled datasets, which they believe produces a better performing model [45]. In experiment 4, an ensemble approach was in fact used that exploited the strengths of both supervised (i.e. trained deep networks) and unsupervised computational methods for image processing. Unsupervised learning allows more salient features and patterns to be identified in pathology images, but requires very large datasets for training purposes. Compared to the first three experiments, the AI solution used in experiment 4 was unique because it also involved CBIR. CBIR systems rely on a search engine to retrieve similar images, and in pathology this has been recognized as an alternate framework to classify images and maximize diagnostic accuracy [46-47]. Google have developed a similar search tool for use in digital pathology called “similar image search for histopathology” (SMILY) [48].

The selection of pathology data used to generate AI algorithms is important, because the quality of the input usually determines the quality of the output. This concept is well conveyed in the computer science acronym GIGO (i.e. *garbage in, garbage out*). There are several pre-imaging variables (e.g. tissue folds, pale staining, blurred areas, and air

bubbles on slides) that can negatively affect the extraction of deep features in images that, in turn, can cause an AI algorithm to perform poorly. For this reason, in experiment 4 whole-slide images of low quality (e.g. slides scanned at low resolution or with large out-of-focus regions) were rejected. It is equally important that data used for training be heterogenous and comprised of a suitable distribution of cases that reflects real-world data likely to be encountered in clinical practice. This is because heterogenous datasets help build more robust algorithms that can be better generalized to work on digital slides from most laboratories. A prior survey involving greater than 500 studies about AI in medical imaging found that 94% only employed data from one site [49]. In an attempt to avoid this pitfall, all of the experiments in this study incorporated data from multiple sources (e.g. UPMC, Wan Fang Hospital, SMC, Maccabi, TCGA). Furthermore, for experiment 2 the datasets included breast carcinoma cases with a broad range of tumor grades and stages. Another drawback inherent in developing AI algorithms is data bias. This was encountered in experiment 1, because most image patches did not contain mycobacteria which resulted in imbalanced datasets (i.e. there were disproportionately many more negative than positive patches). To avoid bias toward a negative prediction, the ratio of negative to positive patches was thus intentionally curbed through random sampling during training.

Developing AI algorithms in Anatomical Pathology shares challenges common to many other deep learning projects. These include lack of readily available datasets, lengthy time needed for humans to label and annotate data, as well as significant investment in computer equipment. However, developing AI algorithms for Anatomical Pathology also

poses unique obstacles. Whole-slide images, for example, are very large files (i.e. several gigabytes in size) and hence often unwieldy to work with. Breaking these digital slides into smaller image patches helped overcome this challenge and also provided a mechanism to augment data for training purposes. For pathologist buy-in to use AI tools it will be important to avoid the “black box” concept. Hence, it is important to deliver results generated by algorithms in an explainable format. For this reason, data output in the experiments conducted was either converted into presentable image galleries or easy to view heatmaps overlaid on H&E images. This appeared to be more palatable for pathologists and easier to use in clinical practice.

The performance of AI algorithms are commonly evaluated using AUC [13]. An AUC value that approaches 1 signifies high discrimination performance. For experiment 1 the AUC = 0.960 at the image patch level and AUC = 0.900 at the digital slide level [18]. For experiment 3 the algorithm developed achieved an AUC of 0.997 for prostate cancer detection in the internal test set and 0.991 in the external validation set [28]. This minor difference in experiment 3 may be attributed to variability in tissue staining, prepared slides used, and the scanned images for the two datasets. Indeed, it is well-known that feature stability between different sites may be compromised by laboratory-specific pre-imaging variables (e.g. specimen fixation, tissue thickness, stain reagents) and the proprietary scanner used to digitize slides [13, 50]. Nevertheless, the AUC for experiment 3 is comparable to other publications on this topic showing that most modern AI models can accurately detect prostate adenocarcinoma with AUC values up to 0.991 to 0.997 [51]. It is important to also recognize that most AI algorithms have

been developed using pathology cases typically encountered in practice, because it is often difficult to collect sufficient numbers of rare cases to train algorithms. Of course, this may negatively impact their performance in real-world practice [52].

Additional testimony about the success of the work reported herein is the ongoing usage for some of the AI tools developed. The AI system used to screen whole-slide images for AFB is still being used daily at UPMC in the USA. The prostate cancer diagnosis AI tool is similarly still being used daily at Maccabi Hospital in Israel, where the workflow was subsequently switched from a second read system (i.e. retrospective case review) to a first read system (i.e. prospective review before a pathologist reviews the case). Finally, the barcode indexing technology used to develop the pathology image search AI tool in experiment 4 has since been adopted to build *Yottixel*, which is another image search engine for large archives of histopathology whole-slide images intended to advance further research in this field [53].

Building AI systems that perform well in clinical practice is challenging and often subject to technical limitations [17,54]. There were indeed several limitations encountered in the various experiments of this study. Whilst the algorithms developed were successful, using larger datasets may have resulted in better performance. The publicly available TCGA dataset that contains thousands of labeled whole-slides with helpful metadata was helpful to address this challenge in experiment 4. The whole-slide images used to train algorithms in this study were only scanned at 40x magnification and using only one focal (z) plane. The reason is that the commercial brightfield whole-slide scanners used

in this study could not scan slides at greater magnification. However, utilizing higher resolution images for training may have generated better algorithm performance. This is especially true for the AI tool used to screen for AFB because it has been shown that scanning slides at 100x with oil immersion is more desirable than 40x for resolving microorganisms in digital images [55]. Some investigators have also recently shown that employing multi-magnification-based machine learning can yield more accurate models [56]. Another limitation in experiment 2 was not standardizing the computer monitors used for annotation and the end-user reader study. Although the impact this may have had is unclear, it has been shown that different types of displays (e.g. medical grade versus consumer off the shelf screens) may influence pathologists' diagnostic performance [57].

The four experiments in this study demonstrate the enormous potential for AI to successfully augment the practice of Anatomical Pathology. However, there is currently limited experience with AI clinical use around the world, and a lack of practical AI validation guidelines for pathologists to follow. Hopefully the published results of my PhD work will provide helpful translational evidence for the future development of such clinical guidelines. Apart from simply producing “weak” AI-based tools that merely replicate the work pathologists perform, it is anticipated that soon there will be “strong” AI systems capable of undertaking challenging tasks pathologists cannot accomplish with their microscopes. However, before these state-of-the-art tools can be expected to be widely adopted, many hurdles (e.g. regulatory approval, financial reimbursement, pathologist buy-in) will need to be overcome. Surmounting these barriers may take

longer in countries such as South Africa where financial resources are limited, which is unfortunate because that is the precise clinical setting where pathologists are scarce, workloads are high, and AI support is likely to be especially beneficial.

1.7 REFERENCES

1. Pantanowitz L. Whole slide imaging, In: Pantanowitz L, Parwani AV, eds. *Digital Pathology*. Chicago: ASCP Press; 2017, pages 59-76.
2. Lara H, Li Z, Abels E, Aeffner F, Bui MM, ElGabry EA, Kozlowski C, et al. Quantitative image analysis for tissue biomarker use: A white paper from the digital pathology association. *Appl Immunohistochem Mol Morphol*. 2021; 29(7):479-493.
3. Evans AJ, Salama ME, Henricks WH, Pantanowitz L. Implementation of Whole Slide Imaging for Clinical Purposes: Issues to Consider From the Perspective of Early Adopters. *Arch Pathol Lab Med*. 2017; 141(7):944-959.
4. Hanna MG, Pantanowitz L. Digital Pathology, In: *Encyclopedia of biomedical engineering*. Narayan R, ed. Philadelphia, Elsevier; 2019, Volume 2, pages 524-532.

5. Evans AJ, Bauer TW, Bui MM, Cornish TC, Duncan H, Glassy EF, et al. US Food and Drug Administration Approval of Whole Slide Imaging for Primary Diagnosis: A Key Milestone Is Reached and New Questions Are Raised. *Arch Pathol Lab Med*. 2018; 142(11):1383-1387.

6. Fraggetta F, Pantanowitz L. Going fully digital: utopia or reality? *Pathologica*. 2018; 110(1):1-2.

7. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019; 25(1):44-56.

8. Louis DN, Gerber GK, Baron JM, Bry L, Dighe AS, Getz G, et al. Computational pathology: an emerging definition. *Arch Pathol Lab Med*. 2014; 138(9):1133-1138.

9. Chang HY, Jung CK, Woo JI, Lee S, Cho J, Kim SW, Kwak TY. Artificial Intelligence in Pathology. *J Pathol Transl Med*. 2019; 53:1–12.

10. Niazi MKK, Parwani AV, Gurcan MN. Digital pathology and artificial intelligence. *Lancet Oncol*. 2019; 20(5):e253-e261.

11. Naugler C, Church DL. Automation and artificial intelligence in the clinical laboratory. *Critical Rev Clin Lab Sciences*. 2019; 56(2):98-110.

12. Yao K, Singh A, Sridhar K, Blau JL, Ohgami RS. Artificial intelligence in pathology: a simple and practical guide. *Adv Anat Pathol*. 2020; 27(6):385-393.
13. Harrison JH, Gilbertson JR, Hanna MG, Olson NH, Seheult JN, Sorace JM et al. Introduction to artificial intelligence and machine learning in pathology. *Arch Pathol Lab Med*. 2021; 145(10):1228-1254.
14. van der Laak J, Litjens G, Ciompi F. Deep learning in histopathology: the path to the clinic. *Nat Med*. 2021; 27(5):775-784.
15. Maddox TM, Rumsfeld JS, Payne PRO. Questions for Artificial Intelligence in Health Care. *JAMA*. 2019; 321(1):31-32.
16. Bi WL, Hosny A, Schabath MB, Giger ML, Birbak NJ, Mehrtash et al. Artificial intelligence in cancer imaging: clinical challenges and applications. *CA Cancer J Clin*. 2019; 69(2):127-157.
17. Cheng J, Balis U, McClintock D, Abel J, Pantanowitz L. Challenges in the development, deployment and regulation of artificial intelligence (AI) in anatomical pathology. *Am J Pathol*. 2021; 191(10):1684-1692.
18. Pantanowitz L, Wu U, Seigh L, LoPresti E, Yeh FC, Salgia P, et al. Artificial intelligence-based screening for mycobacteria in whole-slide images of tissue samples. *Am J Clin Pathol*. 2021; 156(1):117-128.

19. World Health Organization. *Global Tuberculosis Report 2021*. Geneva: World Health Organization; 2021. <https://www.who.int/teams/global-tuberculosis-programme/tb-reports/global-tuberculosis-report-2021> [Last accessed October 23, 2021]

20. Tadrous PJ. Computer-assisted screening of Ziehl-Neelsen-stained tissue for mycobacteria. Algorithm design and preliminary studies on 2,000 images. *Am J Clin Pathol*. 2010; 133(6):849-858.

21. Xiong Y, Ba X, Hou A, Zhang K, Chen L, Li T. Automatic detection of *mycobacterium tuberculosis* using artificial intelligence. *J Thorac Dis*. 2018; 10(3):1936-1940.

22. Wang HS, Liang WY. Automatic *mycobacterium tuberculosis* detection using simple image processing with artificial intelligence. *J Pathol Inform*. 2019; 10:S40.

23. Pantanowitz L, Hartman D, Qi Y, Cho EY, Suh B, Paeng K, et al. Accuracy and efficiency of an artificial intelligence tool when counting breast mitoses. *Diagnostic Pathology* 2020; 51(1):80.

24. Malon C, Brachtel E, Cosatto E, Graf HP, Kurata A, Kuroda M et al. Mitotic figure recognition: agreement among pathologists and computerized detector. *Anal Cell Pathol (Amst)*. 2012; 35(2):97-100.

25. Veta M, van Diest PJ, Willems SM, Wang H, Madabhushi A, Cruz-Roa A et al. Assessment of algorithms for mitosis detection in breast cancer histopathology images. *Med Image Anal.* 2015; 20(1):237-48.
26. Veta M, Heng YJ, Stathonikos N, Bejnordi BE, Beca F, Wollmann T et al. Predicting breast tumor proliferation from whole-slide images: The TUPAC16 challenge. *Med Image Anal.* 2019; 54:111-121.
27. Nateghi R, Danyali H, Helfrroush MS. A deep learning approach for mitosis detection: Application in tumor proliferation prediction from whole slide images. *Artif Intell Med.* 2021; 114:102048.
28. Pantanowitz L, Quiroga-Garza GM, Bien L, Heled R, Laifenfeld D, Linhart C et al. Clinical validation and deployment of an AI-based algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies. *Lancet Digital Health.* 2020; 2(8):e407-e416.
29. Gertych A, Ing N, Ma Z, Fuchs TJ, Salman S, Mohanty S, Bhele S et al. Machine learning approaches to analyze histological images of tissues from radical prostatectomies. *Comput Med Imaging Graph.* 2015; 46(Pt 2):197-208.
30. Bulten W, Balkenhol M, Belinga JJA, Brilhante A, Cakir A, Egevad L et al. Artificial intelligence assistance significantly improves Gleason grading of prostate biopsies by pathologists. *Mod Pathol.* 2021; 34(3):660-671.

31. Van Booven DJ, Kuchakulla M, Pai R, Frech FS, Ramasahayam R, Reddy P et al. A Systematic Review of Artificial Intelligence in Prostate Cancer. *Res Rep Urol*. 2021; 13:31-39.
32. Perincheri S, Levi AW, Celli R, Gershkovich P, Rimm D, Morrow JS et al. An independent assessment of an artificial intelligence system for prostate cancer detection shows strong diagnostic accuracy. *Mod Pathol*. 2021; 34(8):1588-1595.
33. Pantanowitz L, Michelow P, Hazelhurst S, Kalra S, Choi C, Shah S et al. A digital pathology solution to resolve the tissue floater conundrum. *Arch Pathol Lab Med*. 2021; 145(3):359-364.
34. Hunt JL, Swalsky P, Sasatomi E, Niehouse L, Bakker A, Finkelstein SD. A microdissection and molecular genotyping assay to confirm the identity of tissue floaters in paraffin-embedded tissue blocks. *Arch Pathol Lab Med*. 2003; 127:213-217.
35. Venditti M, Hay RW, Kulaga A, Demetrick DJ. Diagnosis of ectopic tissue versus contamination by genetic fingerprinting in a routine surgical pathology specimen. *Hum Pathol*. 2007; 38(2):378-382.

36. Pantanowitz L, Carter A, Kurc T, Sharma A, Sussman A, Saltz J. 20 Years of Digital Pathology: an overview of the road travelled, what is on the horizon, and emergence of vendor neutral archives. *J Pathol Inform.* 2018; 9:46.
37. Chang HY, Jung CK, Woo JI, Lee S, Cho J, Kim SW, Kwak TY. Artificial Intelligence in Pathology. *J Pathol Transl Med.* 2019; 53:1-12.
38. Steiner DF, MacDonald R, Liu Y, Truszkowski P, Hipp JD, Gammage C et al. Impact of Deep Learning Assistance on the Histopathologic Review of Lymph Nodes for Metastatic Breast Cancer. *Am J Surg Pathol.* 2018; 42(12):1636-1646.
39. Benjordi BE, Mullooly M, Pfeiffer RM, Fan S, Vacek PM, Weaver DL et al. Using deep convolutional neural networks to identify and classify tumor-associated stroma in diagnostic breast biopsies. *Mod Pathol.* 2018; 31(10):1502-1512.
40. Hartman DJ, van der Laak JAWM, Gurcan MN, Pantanowitz L. Value of public challenges for the development of pathology deep learning algorithms. *J Path Inform.* 2020; 11:7.
41. Golden JA. Deep learning algorithms for detection of lymph node metastases from breast Cancer: Helping artificial intelligence be seen. *JAMA.* 2017; 318:2184-2186.

42. Sarwar S, Dent A, Faust K, Richer M, Djuric U, Van Ommeran R et al. Physician perspectives on integration of artificial intelligence into diagnostic pathology. *NPJ Digit Med*. 2019; 2:28.

43. Jiang Y, Yang M, Wang S, Li X, Sun Y. Emerging role of deep learning-based artificial intelligence in tumor pathology. *Cancer Commun (Lond)*. 2020; 40(4):154-166.

44. Janowczyk A, Leo P, Rubin MA. Clinical deployment of AI for prostate cancer diagnosis. *Lancet Digit Health*. 2020; 2(8):e383-e384.

45. Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med*. 2019; 25(8):1301-1309.

46. Zheng Y, Jiang Z, Zhang H, Xie F, Ma Y, Shi H et al. Histopathological whole slide image analysis using context-based CBIR. *IEEE Trans Med Imaging*. 2018; 37(7):1641– 1652.

47. Tizhoosh H, Diamandis P, Campbell CJV, Safarpour A, Kalra S, Maleki D et al. Searching Images for Consensus: Can AI Remove Observer Variability in Pathology? *Am J Pathol*. 2021; 191(10):1702-1708.

48. Hegde N, Hipp JD, Liu Y, Emmert-Buck M, Reif E, Smilkov D et al. Similar image search for histopathology: SMILY. *NPJ Digital Medicine*. 2019; 2:56.
49. Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. *Korean J Radiol*. 2019; 20(3): 405-410.
50. Leo P, Lee G, Shih NNC, Elliott R, Feldman MD, Madabuhushi A. Evaluating stability of histomorphometric features across scanner and staining variations: prostate cancer diagnosis from whole slide images. *J Med Imaging*. 2016; 3(4):047502.
51. Egevad L, Delahunt B, Samaratunga H, Tsuzuki T, Yamamoto Y, Yaxley J et al. The emerging role of artificial intelligence in the reporting of prostate pathology. *Pathology*. 2021; 53(5):565-567.
52. Pantanowitz L, Mehra R, Kunju LP. AI reality check when evaluating difficult to grade prostate cancers. *Virchows Arch*. 2021; 478(4):617-618.
53. Kalra S, Tizhoosh H, Choi C, Shah S, Diamandis P, Campbell CJV, Pantanowitz L. *Yottixel* - An image search engine for large archives of histopathology whole slide images. *Medical Image Analysis*. 2020; 65:101757.

54. Tizoosh HR, Pantanowitz L. Artificial intelligence and digital pathology: challenges and opportunities. *J Pathol Inform.* 2018; 9:38.
55. Rhoads DD, Habib-Bein NF, Hariri RS, Hartman DJ, Monaco SE, Lesniak A et al. Comparison of the Diagnostic Utility of Digital Pathology Systems for Telemicrobiology. *J Pathol Inform.* 2016; 7:10.
56. D'Alfonso TM, Ho DJ, Hanna MG, Grabenstetter A, Yarlagaadda DVK, Geneslaw L et al. Multi-magnification-based machine learning as an ancillary tool for the pathologic assessment of shaved margins for breast carcinoma lumpectomy specimens. *Mod Pathol.* 2021; 34(8):1487-1494.
57. Abel JT, Ouillet P, Williams CL, Blau J, Cheng J, Yao K et al. Display Characteristics and Their Impact on Digital Pathology: A Current Review of Pathologists' Future "Microscope". *J Pathol Inform.* 2020; 11:23.

CHAPTER 2: PUBLISHED ARTICLE 1

Pantanowitz L, Wu U, Seigh L, LoPresti E, Yeh FC, Salgia P, Michelow P, Hazelhurst S, Chen WY, Hartman D, Yeh CY. Artificial intelligence-based screening for mycobacteria in whole-slide images of tissue samples. *American Journal of Clinical Pathology*, 2021; 156(1):117-128.

<https://pubmed.ncbi.nlm.nih.gov/33527136/>

Abstract

Objectives: This study aimed to develop and validate a deep learning algorithm to screen digitized acid fast-stained (AFS) slides for mycobacteria within tissue sections.

Methods: A total of 441 whole-slide images (WSIs) of AFS tissue material were used to develop a deep learning algorithm. Regions of interest with possible acid-fast bacilli (AFBs) were displayed in a web-based gallery format alongside corresponding WSIs for pathologist review. Artificial intelligence (AI)-assisted analysis of another 138 AFS slides was compared to manual light microscopy and WSI evaluation without AI support.

Results: Algorithm performance showed an area under the curve of 0.960 at the image patch level. More AI-assisted reviews identified AFBs than manual microscopy or WSI examination ($P < .001$). Sensitivity, negative predictive value, and accuracy were highest for AI-assisted reviews. AI-assisted reviews also had the highest rate of matching the original sign-out diagnosis, were less time-consuming, and were much easier for pathologists to perform ($P < .001$).

Conclusions: This study reports the successful development and clinical validation of an AI-based digital pathology system to screen for AFBs in anatomic pathology material. AI assistance proved to be more sensitive and accurate, took pathologists less time to screen cases, and was easier to use than either manual microscopy or viewing WSIs.

Artificial Intelligence–Based Screening for Mycobacteria in Whole-Slide Images of Tissue Samples

Liron Pantanowitz, MD,^{1,2} Uno Wu, MEng,^{3,4} Lindsey Seigh,¹ Edmund LoPresti,⁵ Fang-Cheng Yeh, PhD,⁶ Payal Salgia, MBBS, DPB, DNB,¹ Pamela Michelow, MBBCh,² Scott Hazelhurst, PhD,⁷ Wei-Yu Chen, MD, PhD,^{8,9} Douglas Hartman, MD,^{1,*} and Chao-Yuan Yeh, MD⁴

From the ¹Department of Pathology and ⁵Information Services Division, University of Pittsburgh Medical Center, Pittsburgh, PA, USA; ²Department of Anatomical Pathology, University of the Witwatersrand and National Health Laboratory Services, Johannesburg, South Africa; ³Department of Electrical Engineering, Molecular Biomedical Informatics Lab, National Cheng Kung University, Tainan City, Taiwan; ⁴aetherAI, Taipei, Taiwan; ⁶Department of Neurological Surgery, University of Pittsburgh, Pittsburgh, PA, USA; ⁷School of Electrical & Information Engineering and Sydney Brenner Institute for Molecular Bioscience, University of the Witwatersrand, Johannesburg, South Africa; and ⁸Department of Pathology, Wan Fang Hospital, and ⁹Department of Pathology, School of Medicine, Taipei Medical University, Taipei, Taiwan.

Key Words: Acid-fast bacilli; Artificial intelligence; Deep learning; Digital pathology; Informatics; Screening; Mycobacteria; Whole-slide imaging

Am J Clin Pathol 2021;XX:1–13

DOI: 10.1093/AJCP/AQAA215

ABSTRACT

Objectives: This study aimed to develop and validate a deep learning algorithm to screen digitized acid fast–stained (AFS) slides for mycobacteria within tissue sections.

Methods: A total of 441 whole-slide images (WSIs) of AFS tissue material were used to develop a deep learning algorithm. Regions of interest with possible acid-fast bacilli (AFBs) were displayed in a web-based gallery format alongside corresponding WSIs for pathologist review. Artificial intelligence (AI)–assisted analysis of another 138 AFS slides was compared to manual light microscopy and WSI evaluation without AI support.

Results: Algorithm performance showed an area under the curve of 0.960 at the image patch level. More AI-assisted reviews identified AFBs than manual microscopy or WSI examination ($P < .001$). Sensitivity, negative predictive value, and accuracy were highest for AI-assisted reviews. AI-assisted reviews also had the highest rate of matching the original sign-out diagnosis, were less time-consuming, and were much easier for pathologists to perform ($P < .001$).

Conclusions: This study reports the successful development and clinical validation of an AI-based digital pathology system to screen for AFBs in anatomic pathology material. AI assistance proved to be more sensitive and accurate, took pathologists less time to screen cases, and was easier to use than either manual microscopy or viewing WSIs.

Key Points

- A deep learning algorithm was developed to screen digitized slides for mycobacteria.
- Detected mycobacteria were displayed in gallery format for review.
- Artificial intelligence assistance was more sensitive, accurate, quicker, and easier to use than manual microscopy or viewing digital slides.

Mycobacteria are a major cause of infectious disease morbidity and mortality worldwide. This includes tuberculosis (TB) caused by the bacillus *Mycobacterium tuberculosis* (MTb). MTb is one of the leading global causes of death, especially in susceptible populations (eg, malnutrition, acquire immune deficiency syndrome) and people living in resource-poor countries. According to the World Health Organization, in 2018, there were around 10 million people ill with TB and 1.2 million TB deaths among human immunodeficiency virus–negative individuals.¹ Most cases were from Southeast Asia, Africa, and the western Pacific. If diagnosed in a timely manner, infected patients may be treated, and their risk of transmitting this communicable disease to others is reduced.

Mycobacteria bacilli are small organisms (length = 2–4 μm ; width = 0.2–0.5 μm) that are consequently hard to detect microscopically without the aid of a special stain. Laboratory tests for TB vary and include microscopic examination of sputum smears, blood tests, culture, and molecular testing. Due to the high content of mycolic acids in their cell walls, mycobacteria can be identified with an

acid-fast stain (AFS). For bright-field microscopy, a Ziehl-Neelsen (ZN) stain is used or modifications such as the Kinyoun and Fite stains. With these stains, rod-shaped bacilli stain red (acid-fast bacilli [AFBs]) and background material blue/green with the counterstain (eg, methylene blue). For fluorescent microscopy, auramine and rhodamine fluorescent stains are used. In situations where TB infection is unsuspected, fresh tissue may not be available for testing. This is often the case with fixed tissue and cytology samples. However, mycobacteria in sections from these fixed specimens are not readily visible with a H&E or a Gram stain. Therefore, to render a microscopic diagnosis, an AFS is often ordered when infection is being considered (eg, granulomas or necrosis are identified) in lung specimens or extrapulmonary sites. This is a simple and relatively cheap method to look for AFBs. While such manual microscopy to screen for AFBs has been reported to have good predictive value,^{2,3} currently this mundane task needs to be performed by trained individuals and is time-consuming (eg, 15-20 minutes/slide is often recommended), laborious, inconsistent, and subject to human error. The sensitivity of manual detection also depends on the number of AFBs present, which can be sparse in early infections.

To try improve the efficiency, raise accuracy, and alleviate busy workloads, attempts have been made to automate AFB microscopy identification by employing digital imaging and computer vision techniques. Jha et al⁴ showed that such automated digital microscopy of sputum smears for detecting MTb could significantly lower the cost of diagnosis, which is important in countries such as South Africa, where resources are low. Panicker et al⁵ evaluated several automatic methods based on image-processing techniques published between 1998 and 2014. Their review article demonstrated that the accuracy of these image algorithms designed to automatically detect AFBs restricted mostly to sputum samples markedly improved over the years. For sputum microscopy, several commercial systems have already been marketed.⁵ However, very few studies to date have similarly looked at applying image analysis for AFB detection in histopathology material using either static snapshots^{6,7} or whole-slide images (WSIs).⁸

The aim of this study was to develop and validate a deep learning algorithm for assisting pathologists with

screening digitized acid fast-stained slides for the detection of mycobacterial infection.

Materials and Methods

Institutional review board approval was obtained for this study (University of Pittsburgh, PRO19060216; University of Witwatersrand, clearance certificate M191003).

Acid-Fast Staining

Slides for this study were contributed by two separate institutions: (1) University of Pittsburgh Medical Center (UPMC), Pittsburgh, PA, and (2) Wan Fang Hospital, Taipei Medical University, Taipei, Taiwan. Sections (0.4 μm thick) were routinely prepared from formalin-fixed, paraffin-embedded pathology material, including surgical pathology samples or cytology cell blocks. At UPMC, acid-fast stains were performed on slides using an Artisan autostainer instrument that employs a commercial AFB stain kit (AR162; Agilent Dako). This automated staining process includes onboard drying and deparaffinization. The AFB stain kit used is a modification of the original ZN method. The AFS process involves the application of carbol fuchsin to stain AFBs red, followed by decolorization of all other tissue elements. Thereafter, a methylene blue counterstain is applied. At Wan Fang Hospital, AFS was performed manually using the Kinyoun (cold) method.

Clinical Data Sets

A total of 441 WSIs were used for algorithm development (Table 1). At UPMC, 297 randomly retrieved archival cases (some with multiple AFS slides) from 2016 to 2019, which were signed out by 26 different pathologists, were included. For each of these cases, the following metadata were recorded: patient sex (55.2% men, 44.8% women), patient age (mean, 55.5 years; range, 19-90 years), sample type (98.7% surgical pathology cases, 1.3% cytology cases), specimen anatomic location (57.2% lung, 42.8% extrapulmonary), original pathology diagnosis, presence of necrosis (present in 19.9% of cases), and AFS interpretation reported by the original sign-out pathologist. As illustrated

Table 1
Breakdown of Whole-Slide Images (WSIs) Used in Acid-Fast Bacilli Algorithm Development

Characteristic	Algorithm Development, No.		Analytical Validation, No.		Testing, No.		Total No.	
	Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative
WSIs	47	371	9	3	6	5	62	379
Patches	4,629	1,049,766	449	40,508	600	21,644	5,678	1,111,918

in Table 1, the slides included a mixture of positive (AFBs present) and negative (AFBs absent) cases. Table 1 also shows how the 441 WSIs were split into three subsets for algorithm development (data set to train our deep neural networks), analytical validation (validation data set to check model performance), and testing (test data set used to calculate and report performance). Wan Fang Hospital in Taiwan contributed an additional 15 positive cases for training purposes derived from lung, pleural, and lymph node tissue. If at least one slide in a case was positive for AFBs, then the overall case was labeled positive. If available, corresponding microbiology results were documented, including culture results (1 positive case, 11 negative cases, not tested in the remaining cases) and AFB polymerase chain reaction (PCR) findings (5 positive cases, 7 negative cases, not ordered in the remaining cases).

A separate randomly selected set of 78 archival cases (138 slides) was used for clinical validation purposes. These cases from UPMC, accessioned between June 2015 and July 2019, were different from the previously used batch of slides and signed out by 18 different anatomic pathologists. The cases used for the test set did not have slides used in the training set. The 78 cases included 55.1% men and 44.9% women, patients ranging in age from 21 to 85 years (mean, 58.5 years), 87.2% surgical cases and 12.8% cytology cases, and 56.4% lung and 43.6% extrapulmonary cases, with necrosis present in 41.0% of cases. Microbiology culture results for mycobacteria included three positive cases, six negative cases, and no test for the remaining cases. In many of these cases, since the diagnosis of tuberculosis was not clinically suspected, cultures were not requested upfront. PCR findings for mycobacteria included 1 positive case and 18 negative cases, and the remaining cases did not have molecular testing ordered.

At UPMC, all slides were scanned at $\times 40$ ($0.25\ \mu\text{m}$ /pixel resolution) with a single Z-focus plane using an Aperio AT2 scanner (Leica Biosystems). The typical resolution for a whole-slide image was $100,000 \times 100,000$ pixels. At Wan Fang Hospital, slides were also scanned with a single-focus plane at $\times 40$ but using a Hamamatsu Nanozoomer XR ($0.23\ \mu\text{m}$ /pixel resolution).

Algorithm Development

We adopted a patch-based approach to the AFB detection problem where a WSI was divided into nonoverlapping patches 64×64 pixels in resolution, and each patch was determined by the algorithm to be positive or negative for AFBs. To achieve a better balance between sensitivity and specificity, the algorithm consisted of two deep convolutional neural network (CNN) models (Figure 1),

in which the first model had higher sensitivity and the second model had high specificity. The reason for using two models was to filter out false-positive predictions. While changing the threshold may have solved this issue, we found that it was not easy to decide on a threshold. The two models have the same architecture, GhostNet,⁹ but were trained differently.

For training data, positive and negative patches were generated by expert annotation on WSIs via a web-based application (aetherSlide). A single AFB occupied a minute component (10×10 resolution) of an entire WSI ($100,000 \times 100,000$ resolution). For positive annotations (AFBs identified; $n = 6,817$), a small square region of the image containing organism(s) was cropped. AFBs were permitted to be present anywhere within the region of a positive patch. Negative patches consisted of larger polygonal regions, and over one million patches without AFBs were sampled. Negative annotations included three different labels: true negative ($n = 2,773$), hard negative areas with AFB mimics, including artifacts and stain precipitate ($n = 7,426$); and background regions ($n = 828$). When training the neural network model, true and hard negative annotations were merged. There were disproportionately (>200 times) more negative patches than positive ones. To prevent neural network models from an inherent bias toward a negative prediction, we limited the ratio of negative to positive patches to 7:3 through random sampling during training. We found that this sampling strategy yielded better algorithm performance over a focal loss strategy.

The first model was pretrained on the CIFAR-10 data set (50,000 images with 32×32 pixels)¹⁰ and further trained using the entire training set. For each

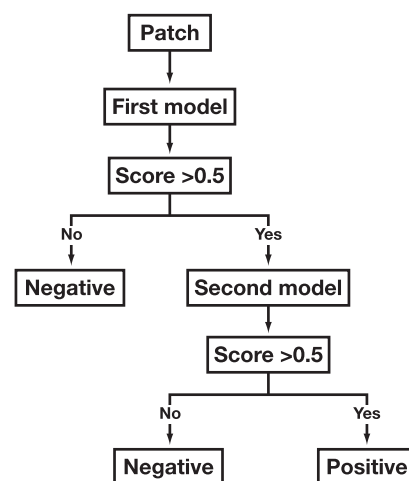


Figure 1 Schematic showing acid-fast bacilli detection algorithm training workflow.

image, we applied random shift, random rotation (0, 90, 180, or 270 degrees), scaling, changing brightness, and horizontal flipping. Pixel values were divided by 127.5 and minus 1. The parameter of L2 regularization was 0.0005, and batch size was 128. We used a stochastic gradient descent (SGD) optimizer with an initial learning rate of 0.1 and momentum of 0.9. The value of the learning rate was divided by 10 when the epoch came to 100, 200, and 300. In total, we trained 400 epochs. The second model was initialized with the weights of the trained first model and fine-tuned using only patches predicted as positive by the first model. The method used to finetune the first and second models was the same. For the second model, we finetuned all layers in the models without L2 regularization. In this model, batch size was 32. We again used the SGD optimizer with an initial learning rate of 0.001 and momentum of 0.9. The learning rate was multiplied 0.98 after every 500 steps. In total, we trained 50,000 steps to avoid overfitting. When calculating validation performance, we used all positive patches and 2,000 random negative patches from the validation set to calculate validation accuracy. We subsequently saved the model with the highest validation accuracy. Color normalization was not employed.

The hardware environment for algorithm development included two central processing units (Intel Xeon CPU E5-2697 v3 at 2.60 GHz, with 14 cores per processor), had a graphics processing unit (Nvidia GeForce

RTX 2080ti), and required random access memory (DDR4-32G * 24) of 768 GB. The software environment included the Ubuntu 18.04 operating system and Python 3.6 programming language. Depending on the file size, it takes the algorithm 5 to 30 minutes to process an entire WSI. This complete process includes reading a WSI, splitting the WSI into patches, and running the machine learning algorithm. Although the program processes one WSI at a time, all of the patches are processed in parallel at the same time. While one CPU with eight cores and 32 GB of RAM is likely sufficient to process a WSI, the Nvidia GPU helps speed up the algorithm.

Web Portal

For managing WSIs in routine clinical practice, a slide processing assistant was developed in C# that identified the digitized slide submitted for processing, made the WSI available to the CNN for analysis, and stored the algorithm's results in a structured query language database. The algorithm's output was transferred to a comma-separated values file, including the (1) X and Y coordinates and (2) likelihood confidence score of predicted positive patches. To make the algorithm output explainable and easy for pathology end users, patches (ie, regions of interest) were displayed via a web portal in gallery format (Figure 2). The web portal was written in C# and JavaScript.

The web-based gallery includes image patches shown as thumbnails, ranked from highest to lowest based on

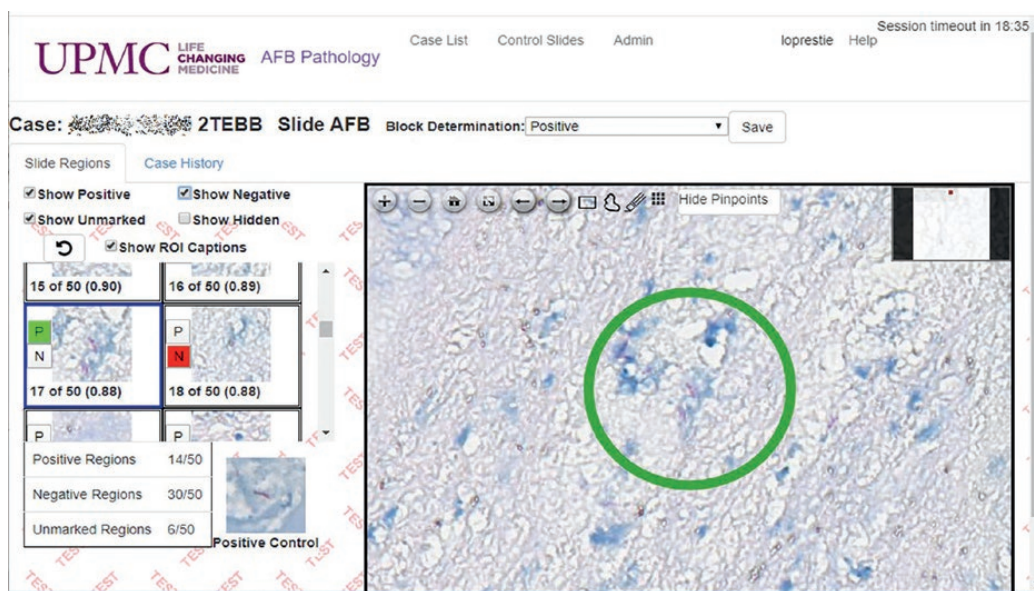


Figure 2 Screenshot from the web portal showing regions of interest (patches) identified by the algorithm in the gallery on the left and the corresponding whole-slide image (WSI) on the right. In this example, regions 17 and 18 (out of 50 total) identified by the algorithm each had a confidence score of 0.88. Clicking on thumbnail 17 displayed acid-fast bacilli (green circle) within the WSI at the exact location where they were detected.

their probability of containing AFBs. These thumbnail images are dynamically generated using OpenSlide¹¹ by automatically taking 80×80 (resolution 96 pixels/inch) snapshots (file size 2 KB) at maximum magnification, centered on the X and Y location within the WSI. A user can scroll down the webpage to examine all the regions of interest for an analyzed digital slide. Each thumbnail also shows the probability (eg, score 0.5 indicates that the image patch has a 50% probability of containing an AFB). For each case, the website also displays the corresponding WSI. Clicking on a thumbnail in the gallery displays that exact region of interest in the context of the WSI image on the right, allowing a pathologist to visualize the AFBs within the context of the entire slide (eg, within a Langhans cell or located extracellularly in a necrotic background). Users are free to navigate through the entire WSI by panning and zooming. They can also define boundaries within the WSI to inspect, filtering out regions outside the boundary (ie, only patches within the boundary get displayed in the gallery).

Clinical Validation

The algorithm and workflow incorporating the web portal to display artificial intelligence (AI)-based results were validated at one site (UPMC) in a blinded study by comparing manual bright-field light microscopy and WSI evaluation of AFS slides to AI-assisted screening independently by two pathologists (pathologists A and B). There was a 7-day minimum washout period between review methods. This washout was selected for logistic reasons. While 2 weeks is often used for digital pathology validation studies, this is debated with arguments for selecting both longer and shorter washout periods. AI-based reviews also occurred last for logistic purposes. For manual microscopy, traditional light microscopes (Olympus BX45 or BX46) were used. The WSIs were examined using 24-inch monitors (Dell or HP) in landscape orientation, each with $1,920 \times 1,080$ resolution and standard dynamic range color space. When reviewing cases with AI support via the web portal, a minimum of 300 thumbnails/case was evaluated even though some cases had many more regions of interest displayed in the gallery. The reviewing pathologists recorded the range of time their assessments took (0-5 minutes, 6-10 minutes, 11+ minutes), AFB interpretation (negative, indeterminate, positive), quantity of AFBs identified (2 or fewer, 3-10, 11+ AFBs), and task difficulty (easy, medium, hard) for each slide. For reviews with the aid of the algorithm, the web portal recorded the precise review time in milliseconds, number of positive regions the review pathologist detected, and the total number of regions recommended

by the algorithm for review. All interpretations using all three of these modalities were compared with the original sign-out diagnoses and, in a subset of these cases, also to available microbiology (culture/PCR) results. Discrepant cases were not adjudicated.

Statistical Analysis

A receiver operating characteristic (ROC) curve was plotted and area under the ROC curve (AUC) was calculated to determine the accuracy of the algorithm once the training phase was completed using Python (scikit-learn¹² and matplotlib¹³). For the clinical validation study, pathologists' interpretations were compared with each other, as well as with the original sign-out interpretation. All 138 slides were compared for both pathologists for each review method (microscope, WSI, algorithm assisted), including AFB interpretation, time to review, and difficulty. Comparisons to the original sign-out and original microscopic findings were based on the 78 cases. If any slide in a multislide case was positive, the case was considered positive. Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy were calculated for each review method, using the signed-out assessment as the gold standard. Interpretations were also compared with available microbiology results. PCR and other microscopic results were either not performed or not available for most cases used in the validation phase of the study. Therefore, results of microscopic findings were not separately used as a gold standard for sensitivity, specificity, NPV, PPV, and accuracy calculations. However, positive results would have affected the AFB assessment at sign-out and therefore are indirectly related to the gold-standard comparison used in this study (sign-out interpretation).

The Pearson χ^2 test was used to determine if the proportion of cases positive for AFBs was different among the review methods (microscope, WSI, AI) compared with the original interpretation. The Pearson χ^2 test was also used to determine if the time category, categorized quantity of AFBs, or difficulty in assessment was different among the reviewers. The normality of recorded distributions for continuous variables (AI-recorded review time, number of positive regions, percentage of total regions that were positive) was examined using the Shapiro-Wilk normality test. As the data were not normally distributed, nonparametric statistical tests were used. The Mann-Whitney test was used to compare pathologists' review time (minutes) and, when using AI-assisted review, the positive regions identified. Statistical significance was assumed

at $P < .05$. Analyses were performed using IBM SPSS Statistics 22 (SPSS).

Results

Algorithm Performance

The final AI-based algorithm was able to reliably detect AFBs in WSIs **Figure 3** and **Figure 4**, even when these microorganisms were not always in clear focus due to the limited image acquisition resolution of the commercial WSI scanners used. **Table 2** summarizes the performance of the algorithm at the image patch and WSI levels. The final

AUC was 0.960 at the image patch level **Figure 5** and 0.900 at the WSI level. To calculate the WSI-level AUC score, a predicted score was computed for each image via averaging all predicted scores of patches higher than 0.5. If a WSI had no patch with a score over 0.5, the WSI-level predicted score was 0. **Table 3** shows the algorithm performance at the patch level if we use only the first, only the second, or both models. This illustrates that using both models provided the highest AUC. In rare cases with abundant AFBs present, the algorithm returned many positive regions. Since too many image patches delayed the web-based application as they took long to upload into the gallery, a forced cutoff of 1,000 images/slide was implemented.

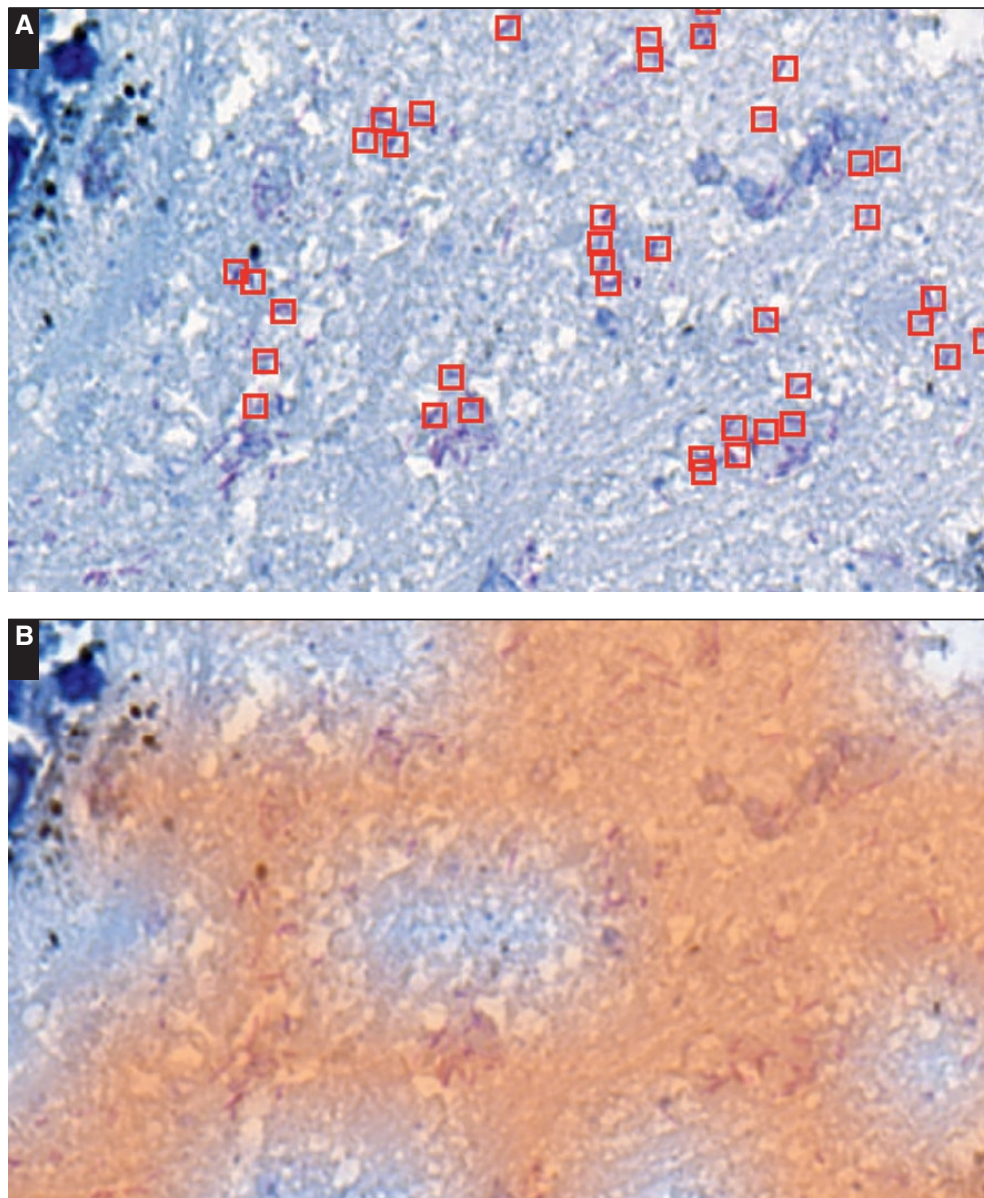


Figure 3 Artificial intelligence–assisted detection of acid-fast bacilli (AFBs) is shown in a whole-slide image (WSI). **A**, Portion of WSI is shown with expert manual annotation (red squares) of AFBs. **B**, Predicted heatmap (orange) from the algorithm is shown overlaid on the exact same region of the WSI.

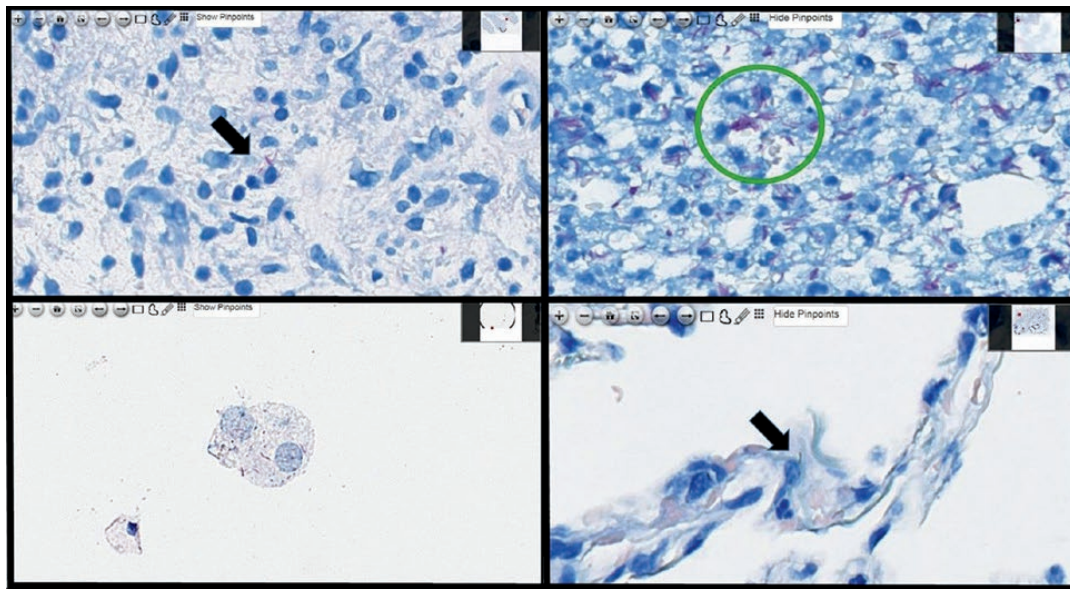


Figure 4 Examples of artificial intelligence–detected acid-fast bacilli (AFB) structures. Upper left: sparse extracellular (arrow) AFBs (algorithm probability 0.98). Upper right: multiple AFBs within the green circle (algorithm probability 1.00). Lower left: scant intracellular AFBs in a cytology specimen (algorithm probability 0.94). Lower right: AFB mimic (arrow) located within normal tissue (algorithm probability 0.54).

Table 2
Algorithm Performance at Image Patch and WSI Levels

Image	Sensitivity	Specificity	AUC
Patch level	0.600	0.999	0.960
WSI level	0.833	0.800	0.900

AUC, area under the curve; WSI, whole-slide image.

Clinical Validation Findings

A statistically significantly higher proportion of AI-assisted slide reviews identified the presence of AFBs (20.3%) than manually using a microscope (11.6%) or by the WSI (7.6%) review method ($\chi^2 = 7.787$, $P = .005$; $\chi^2 = 18.488$, $P < .001$, respectively). **Figure 6** shows the overall proportion of slides identified as positive, as well as the estimation of the quantity of AFBs. AFBs were least likely to be detected when pathologists screened slides using WSI. Pathologist B indicated a significantly higher proportion of positive WSI slides (10.9%) than pathologist A (4.3%) ($\chi^2 = 4.175$, $P = .041$). However, there was not a significant difference in the proportion of positive cases identified by either reviewer using microscope and algorithm-assisted methods. The number of thumbnails identified as positive was significantly higher for pathologist B (median, 0.0%; mean rank, 145.2) than pathologist A (median, 0.0%; mean rank, 131.8) ($U = 8,601.5$, $P = .040$). Algorithm-assisted reviews had a higher rate of matching the original sign-out assessment (84.6%) than did reviews using the microscope (82.7%) and WSI

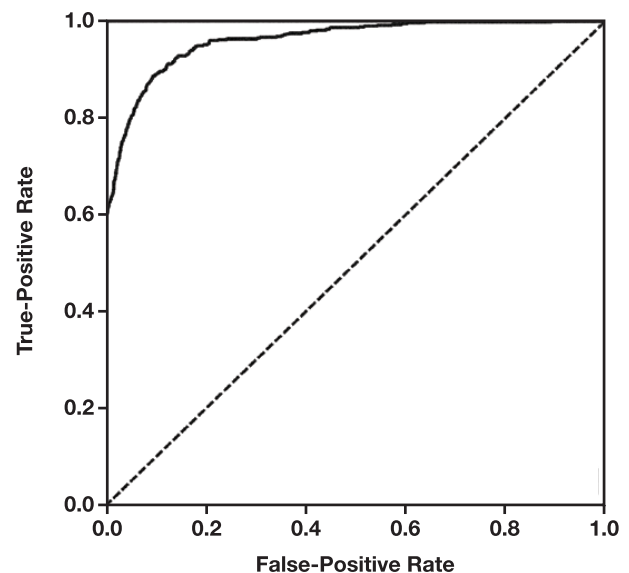


Figure 5 Area under the curve (AUC) for acid-fast bacilli algorithm detection in image patches. AUC = 0.960.

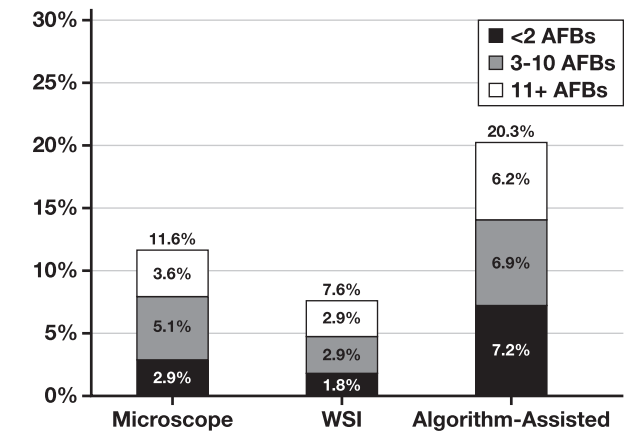
Table 3
Comparison of Algorithm Performance Using Individual or Combined Convolutional Neural Network Models

Applying CNN Models	Specificity	Sensitivity	AUC
First model only	0.96	0.66	0.95
Second model only	1.00	0.61	0.92
Both models	1.00	0.60	0.96

AUC, area under the curve; CNN, convolutional neural network.

(77.6%), although the differences were not statistically significant ($\chi^2 = 0.211$, $P = .646$; $\chi^2 = 2.529$, $P = .112$, respectively). ■Table 4■ shows the sensitivity, specificity, PPV, NPV, and accuracy calculations by review method and reviewer per case. Since the microscope and WSI methods had no false positives, the specificity and PPV for both of these review methods were 100%. The sensitivity, NPV, and accuracy were highest for algorithm-assisted reviews.

Only limited cases had follow-up microbiology results available for comparison. For the three cases with subsequent positive microbiology cultures, only two were interpreted as positive with all review modalities (one slide was missed by one pathologist using manual microscopy). The case with a positive PCR result had four slides that were included. Pathologist A provided a negative AFB assessment for all four slides for each review method. Pathologist B provided a positive result for one of the four slides on microscope and WSI review but a negative result for all four slides on algorithm-assisted review.



■Figure 6■ Proportion of slides positive for acid-fast bacilli (AFBs) screened by different modalities. WSI, whole-slide image.

■Table 4■
Sensitivity, Specificity, PPV, PPV, and Accuracy by Reviewer and Method^a

Characteristic	Microscope Review			Review Using WSI			Algorithm Assisted		
	Pathologist A	Pathologist B	Overall	Pathologist A	Pathologist B	Overall	Pathologist A	Pathologist B	Overall
TP, No.	11	16	27	6	13	19	17	18	35
TN, No.	51	51	102	51	51	102	49	48	97
FP, No.	0	0	0	0	0	0	2	3	5
FN, No.	16	11	27	21	14	35	10	9	19
Sensitivity, %	40.7	59.3	50.0	22.2	48.1	35.2	63.0	66.7	64.8
Specificity, %	100.0	100.0	100.0	100.0	100.0	100.0	96.1	94.1	95.1
PPV, %	100.0	100.0	100.0	100.0	100.0	100.0	89.5	85.7	87.5
NPV, %	76.1	82.3	79.1	70.8	78.5	74.5	83.1	84.2	83.6
Accuracy, %	79.5	85.9	82.7	73.1	82.1	77.6	84.6	84.6	84.6

FN, false negative; FP, false positive; NPV, negative predictive value; PPV, positive predictive value; TN, true negative; TP, true positive; WSI, whole-slide image.
^aSensitivity = TP / (TP + FN). Specificity = TN / (TN + FP). PPV = TP / (TP + FP). NPV = TN / (TN + FN). Accuracy = (TP + TN) / (TP + TN + FP + FN).

There was a statistically significant relationship between review time and pathologist for each review method ■Figure 7■. A statistically significantly higher proportion of algorithm-assisted reviews took 0 to 5 minutes (91.7%) than did microscope (66.3%) or WSI (47.1%) reviews ($\chi^2 = 53.480$, $P < .001$; $\chi^2 = 129.022$, $P < .001$, respectively). ■Table 5■ shows the amount of time for pathologist review by method. A significantly higher proportion of pathologist A's microscopic reviews were completed in 0 to 5 minutes (89.1%) compared with pathologist B (43.5%) ($\chi^2 = 64.366$, $P < .001$). Pathologist A was also significantly quicker at completing their WSI reviews within 0 to 5 minutes (89.1%) compared with pathologist B (5.1%) ($\chi^2 = 195.672$, $P < .001$). However, for algorithm-assisted reviews, a significantly higher proportion of pathologist B's reviews were completed in 0 to 5 minutes (99.3%) compared with pathologist A (84.1%) ($\chi^2 = 20.917$, $P < .001$). The Mann-Whitney results showed that with AI support, pathologist B spent a median of 1.1 min/slide (mean rank, 106.5) compared with pathologist A, who spent a median of 2.5 min/slide (mean rank, 170.5) ($U = 5,107.5$, $P < .001$). Slides with positive regions identified took significantly longer to review (median, 3.6 minutes; mean rank, 180.9) than slides with no positive regions (median, 1.5 minutes; mean rank, 128.9) ($U = 3,574.5$, $P < .001$).

A statistically significantly higher proportion of algorithm-assisted reviews were classified by pathologists as easy (93.5%) compared with manually screening slides with a microscope (43.8%) or using WSI (38.4%) ($\chi^2 = 158.014$, $P < .001$; $\chi^2 = 186.366$, $P < .001$, respectively). Most of the algorithm-assisted reviews were considered easy by both pathologists (92.0% for pathologist A, 94.9% for pathologist B). ■Figure 8■ shows the overall proportion of difficulty level by review method. A significantly higher proportion of microscope reviews were classified as hard (19.9%) compared with algorithm-assisted

reviews (0.0%) ($\chi^2 = 61.087$, $P < .001$). A significantly higher proportion of WSI reviews were classified as medium difficulty (24.3%) compared with algorithm-assisted reviews (6.5%) ($\chi^2 = 33.388$, $P < .001$).

Discussion

This article reports the successful development and clinical validation of a deep learning algorithm for screening digitized AFS slides for the detection of mycobacteria. This AI-based method proved to be more sensitive and accurate, took pathologists less time to screen cases, and was easier to use than either manual microscopy or viewing WSIs. The algorithm was tuned to have high specificity at the expense of sensitivity because the proportion of positive to negative cases in our data set was around 1:226. Hence, even a slight decrease in

specificity score will result in a large unwanted increase in the number of false-positive predictions. Furthermore, doing an AFS on a tissue sample in which TB was not clinically suspected is not really a true screening test. For this use case in anatomic pathology, specificity is thus probably more important to reliably detect AFBs. For the clinical validation, a 7-day minimum washout period was used between review methods. We acknowledge that this is a limitation of this study as there may be recall bias.¹⁴ Perhaps using a different washout period (eg, greater than 4 weeks) or even order in which test modalities were evaluated may have influenced the results. As anticipated, humans with AI assistance do a better job than humans alone at seeking out minute microorganisms (rare events) on a glass or digital slide. For screening purposes, it was important for us to have a sensitive AI tool. Somoskövi et al¹⁵ showed that for manual acid-fast microscopy, examining fewer than 300 fields of view may decrease sensitivity. However, with current WSI technology where image navigation typically requires a computer mouse, humans are unlikely to be able to reliably inspect every field of view in a WSI looking for AFBs at high magnification. An AI system, on the other hand, can easily analyze all of the pixels in a WSI for appropriately stained bacilli. The algorithm we developed was able to detect AFBs whether they were sparse or present in abundance. The novel contribution of our study is that we were able to convert the data output of our AI algorithm (ie, detection of AFBs within image patches) to an explainable format (ie, image gallery displaying image patches ranked based on their likelihood of containing AFBs) that makes it more palatable for pathologists to use and hence deploy in routine clinical practice.

Going “fully digital” (ie, rendering primary diagnoses using WSIs instead of glass slides) without AI support for certain tasks is clearly insufficient because even when performing the mundane task of screening for AFBs, these data

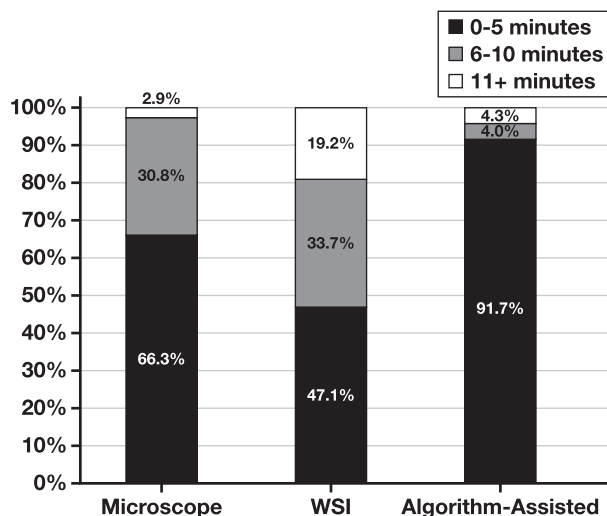


Figure 7 Review time by screening modality. WSI, whole-slide image.

Table 5
Comparison of Time Category for Pathologists by Review Method (n = 138)

AFB Screening Modality	Pathologist A, No. (%)	Pathologist B, No. (%)	χ^2	P Value
Microscope review				
0-5 minutes	123 (89.1)	60 (43.5)	64.366	<.001
6-10 minutes	15 (10.9)	70 (50.7)	51.426	<.001
11+ minutes	0 (0.0)	8 (5.8)	8.239	.004
Review using WSI				
0-5 minutes	123 (89.1)	7 (5.1)	195.672	<.001
6-10 minutes	13 (9.4)	80 (58.0)	72.799	<.001
11+ minutes	2 (1.4)	51 (37.0)	56.069	<.001
Algorithm assisted				
0-5 minutes	116 (84.1)	137 (99.3)	20.917	<.001
6-10 minutes	10 (7.2)	1 (0.7)	7.669	.006
11+ minutes	12 (8.7)	0 (0.0)	12.545	<.001

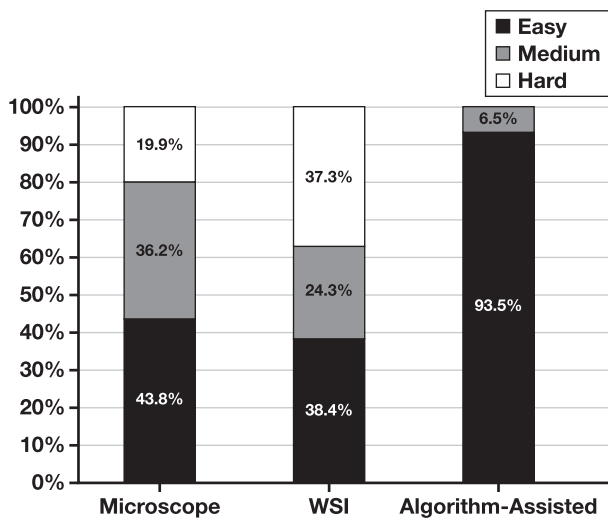


Figure 8 Pathologist-perceived difficulty of identifying acid-fast bacilli by screening modality. WSI, whole-slide image.

show that viewing WSIs was less accurate, took longer, and was harder to do than manual microscopy. The proposed workflow in our study offers pathology laboratories a semiautomatic computer-assisted system to search for AFBs. Even though slide digitization was coupled with AI-based analysis, pathologist oversight was still required to supply a final interpretation. A combination of several components contributed to the overall clinical performance of our AI-based screening method. These include using a deep learning algorithm, ranking retrieved regions of interest according to their likelihood of containing AFBs in a web-based gallery format, and coupling this display with the ability for pathologists to explore corresponding WSIs. Having the WSI accessible to navigate each case was very valuable. This allowed the pathologist to distinguish true pathogenic bacilli (eg, AFBs associated with granulomas and caseous necrosis) from mimics, contaminants, or artifacts.

In rare cases, artifacts that mimicked AFBs were sometimes erroneously identified, but these were easily discernable by pathologists and justifies keeping humans involved in the review process. We did not track the quantity or subtypes of these artifacts. The algorithm we employed was also not trained to detect all of the different species of mycobacteria. While most AFBs are typically rod shaped, they may occasionally be slightly curved, beaded with nonuniform staining, or can stick together to form long cords. This can be overcome in the future by expanding training data sets to include such heterogeneous cases. Although we scanned glass slides at the maximum of $\times 40$ offered by the commercial scanners used in this study, WSI resolution and sometimes out-of-focus areas were limiting factors when trying to reliably

visualize AFBs. Hanna and Pantanowitz¹⁶ previously showed the inability of various WSI scanners to reliably resolve microorganisms such as AFBs. Not surprisingly, for telemicrobiology purposes, WSI scanning at $\times 100$ oil magnification has been recommended.¹⁷ The challenge of focusing on bacilli when specifically designing automated systems to detect AFBs has been recognized by others.^{8,18}

Color plays an important role in the accurate diagnosis of WSI,¹⁹ especially with regard to infectious diseases.²⁰ Prior researchers working on image analysis solutions to detect AFBs were aware of the importance of this feature and had problems related to complex color variations in their figures, as well as color-based pixel segmentation and classification. Costa et al,²¹ for example, attempted to overcome this drawback by designing their AFB algorithm using red minus green images from the RGB color format for bacilli segmentation. We opted not to apply special calibration techniques to WSI-generated images, use color normalization as suggested by some investigators,²² or rely on special color filters in our AI model. Instead, we trusted the CNN to reliably identify red-stained AFBs from background material. Kuok et al²³ from Taiwan similarly found better performance when using a CNN without any preprocessing mechanism for their AFB algorithm. Furthermore, incorporating AFS slides from different institutions and using slides for training from different time periods with stain variation, instead of a single batch stained set of slides, likely contributed to a more robust algorithm.

Table 6 summarizes prior attempts reported in the literature over the past two decades leveraging image analysis to identify mycobacteria.^{6,8,23-32} This comprehensive table does not include all projects in the field, such as one student's related master's thesis from the University of Cape Town in South Africa.³³ Most researchers have developed computer-based systems to automate the detection of AFBs in sputum smears with fluorescent or bright-field microscopy. This is not surprising, because in high-prevalence countries, sputum analysis is the most commonly ordered laboratory test to diagnose MTb. Some automated sputum smear microscopy systems have even been commercialized, such as the "TBDx system" evaluated by Lewis et al²⁹ from South Africa. Unlike tissue, sputum from a patient infected with MTb usually contains abundant AFBs. The lower density of AFBs in tissue specimens thus makes the automated detection of AFBs in tissue sections more difficult. To our knowledge, we are aware of only three other studies in which investigators attempted to automate AFB detection in tissue samples.^{6,8,32} To date, only a few other groups have also successfully shown the ability to screen WSIs for AFBs.^{8,31,32} Our study demonstrates the success of using an AI-based system to screen for AFBs in WSIs of AFS tissue sections. We also used larger training and validation

Table 6
Summary of Published Studies Using Image Analysis to Identify Mycobacteria

Reference	Year	Samples	Stain	Images	Magnification	Performance
Veropoulos et al ²⁴	1999	Sputum smears	Auramine	Snapshots	Not reported	Sensitivity 94.1%
Forero et al ²⁵	2006	Sputum and nonrespiratory samples	Auramine	Snapshot	×25	Sensitivity 97.89%; specificity 94.67%
Sadaphal et al ²⁶	2008	Sputum smears	ZN stain	Snapshots	Not reported	Not reported
Khutlang et al ²⁷	2010	Sputum smears	ZN stain	Snapshots	×100 oil	Not reported
Osibote et al ²⁸	2010	Sputum smears	ZN stain	Snapshots	Not reported	Not reported
Lewis et al ²⁹	2012	Sputum smears	Auramine	Snapshots	×40	Sensitivity 75.8%; specificity 43.5%
Chang et al ³⁰	2012	Sputum smears	Auramine	Snapshots	×200-300	Precision 89.2%
Zingue et al ³¹	2018	Sputum smears	Kinyoun stain	WSI	×100 oil	Sensitivity 97.06%; specificity 86.44%
Kuok et al ²³	2019	Sputum smears	Not reported	Snapshots	×40	Sensitivity 97.67% ^a ; specificity 91.02% ^a
Tadrous ⁶	2010	Tissue	ZN stain	Snapshots	×60	Not reported
Xiong et al ⁸	2018	Tissue	Not reported	WSI	×40	Sensitivity 97.94%; specificity 83.65%
Wang and Liang ³²	2019	Tissue	Not reported	WSI	Not reported	Accuracy 94%

WSI, whole-slide image; ZN, Ziehl-Neelsen.

^aPerformance using a deep convolutional neural network method.

data sets compared with others. For example, Xiong et al⁸ from China used 230 samples for their deep learning model. In addition, Wang and Liang³² used 50 positive samples for training and another 50 WSIs for validation. Tavolara et al³⁴ developed a deep learning algorithm using a database of 178 WSIs to segment MTb bacilli clusters from AFS lung biopsy specimens obtained from experimentally infected mice. We implemented their algorithm on our trained data set and found a very similar performance (for specificity: our algorithm = 0.999 vs Tavolara et al³⁴ = 0.996; for sensitivity: our algorithm = 0.600 vs Tavolara et al³⁴ = 0.598). As their algorithm did not predict a score for each patch, instead of comparing AUC scores, we thus measured the F1 score to compare performances (our algorithm = 0.746 vs Tavolara et al³⁴ = 0.687). The F1 score is an average score of precision (positive predictive value) and recall (sensitivity). Of note, albeit very few cytology cases were included in our study, we were able to demonstrate the feasibility of our algorithm to analyze digitized AFS slides of cell block sections. Although an additional 15 positive cases were used for training purposes from Wan Fang Hospital in Taiwan, future work will require using much larger external validation data sets to assess the generalizability of this algorithm.

In summary, this study reports the effective development and clinical validation of an AI-based digital pathology system to find AFBs in tissue and cytology material. In clinical practice, all that pathologists often require to diagnose a case as positive for TB is the presence of at least one mycobacterium within tissue that reliably stains with an AFS. Hence, detecting just one patch from an image that contains AFBs can help pathologists with clinical decision

making. In addition, given the wide variability in practice regarding how pathologists approach AFB interpretation,³⁵ such an AI-based digital pathology system could potentially improve standardization. Such a system has great potential to expedite the diagnosis of TB, reduce pathologist workload, improve diagnostic accuracy, and even quantitate AFBs if needed. While negative cases could be reported without human review after AI-based screening, the proposed workflow with our setup requires a pathologist to confirm the final diagnosis on all cases. In the near future, many AI-based systems are anticipated to be adopted in clinical microbiology.^{36,37} Indeed, in a review article, Peiffer-Smadja et al³⁸ already found 97 machine learning systems targeting bacterial, viral, parasitic, and fungal infections, all designed to assist clinical microbiologists. We feel that this technology will be a welcome addition to the practice of pathology.

Corresponding author: Liron Pantanowitz MD; lpantanowitz@gmail.com.

Acknowledgments: We thank Colleen Vrbic from Analytical Insights for her help with our statistical analysis and Chitra Sharma and Cynthia Carbine for their administrative help.

Liron Pantanowitz is a consultant for Hamamatsu and on the medical advisory board for Ibex. Uno Wu and Joe Yeh are employees of aetherAI. Every effort was made to avoid a conflict of interest that could potentially influence the study conclusions.

References

1. World Health Organization. Global tuberculosis report 2019. <https://apps.who.int/iris/bitstream/handle/10665/329368/9789241565714-eng.pdf?ua=1>. Accessed June 8, 2020.

2. Murray PR, Elmore C, Krogstad DJ. The acid-fast stain: a specific and predictive test for mycobacterial disease. *Ann Intern Med.* 1980;92:512-513.
3. Lipsky BA, Gates J, Tenover FC, et al. Factors affecting the clinical value of microscopy for acid-fast bacilli. *Rev Infect Dis.* 1984;6:214-222.
4. Jha S, Ismail N, Clark D, et al. Cost-effectiveness of automated digital microscopy for diagnosis of active tuberculosis. *PLoS One.* 2016;11:e0157554.
5. Panicker RO, Soman B, Saini G, et al. A review of automatic methods based on image processing techniques for tuberculosis detection from microscopic sputum smear images. *J Med Syst.* 2016;40:17.
6. Tadrous PJ. Computer-assisted screening of Ziehl-Neelsen-stained tissue for mycobacteria: algorithm design and preliminary studies on 2,000 images. *Am J Clin Pathol.* 2010;133:849-858.
7. Osman MK, Mashor M, Jafaar H. Performance comparison of clustering and thresholding algorithms for tuberculosis bacilli segmentation. *Aust J Basic Appl Sci.* 2011;5:1270-1279.
8. Xiong Y, Ba X, Hou A, et al. Automatic detection of mycobacterium tuberculosis using artificial intelligence. *J Thorac Dis.* 2018;10:1936-1940.
9. Han K, Wang Y, Qi T, et al. GhostNet: more features from cheap operations. https://openaccess.thecvf.com/content_CVPR_2020/papers/Han_GhostNet_More_Features_From_Cheap_Operations_CVPR_2020_paper.pdf.
10. Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images. 2009. <https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf>. Accessed June 8, 2020.
11. Goode A, Gilbert B, Harkes J, et al. OpenSlide: a vendor-neutral software foundation for digital pathology. *J Pathol Inform.* 2013;4:27.
12. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res.* 2011;12:2825-2830.
13. Hunter JD. Matplotlib: a 2D graphics environment. *Comput Sci Eng.* 2007;9:90-95.
14. Campbell WS, Talmon GA, Foster KW, et al. Visual memory effects on intraoperator study design: determining a minimum time gap between case reviews to reduce recall bias. *Am J Clin Pathol.* 2015;143:412-418.
15. Somoskövi A, Hotaling JE, Fitzgerald M, et al. Lessons from a proficiency testing event for acid-fast microscopy. *Chest.* 2001;120:250-257.
16. Hanna M, Pantanowitz L. Comparison of different digital image resolutions for identifying acid fast bacilli. *J Pathol Inform.* 2017;8:S43-S44.
17. Rhoads DD, Habib-Bein NF, Hariri RS, et al. Comparison of the diagnostic utility of digital pathology systems for telemicrobiology. *J Pathol Inform.* 2016;7:10.
18. Mateos-Pérez JM, Redondo R, Nava R, et al. Comparative evaluation of autofocus algorithms for a real-time system for automatic detection of *Mycobacterium tuberculosis*. *Cytometry A.* 2012;81:213-221.
19. Clarke EL, Treanor D. Colour in digital pathology: a review. *Histopathology.* 2017;70:153-163.
20. Salmon R. Color truth for infectious disease imaging. *Pathologist.* 2020. <https://thepathologist.com/inside-the-lab/color-truth-for-infectious-disease-imaging>. Accessed June 2, 2020.
21. Costa MG, Costa Filho CF, Sena JF, et al. Automatic identification of *Mycobacterium tuberculosis* with conventional light microscopy. *Annu Int Conf IEEE Eng Med Biol Soc.* 2008;2008:382-385.
22. Roy S, Kumar Jain A, Lal S, et al. A study about color normalization methods for histopathology images. *Micron.* 2018;114:42-61.
23. Kuok CP, Horng MH, Liao YM, et al. An effective and accurate identification system of *Mycobacterium tuberculosis* using convolution neural networks. *Microsc Res Tech.* 2019;82:709-719.
24. Veropoulos K, Learmonth G, Campbell C, et al. Automated identification of tubercle bacilli in sputum: a preliminary investigation. *Anal Quant Cytol Histol.* 1999;21:277-282.
25. Forero MG, Cristóbal G, Desco M. Automatic identification of *Mycobacterium tuberculosis* by Gaussian mixture models. *J Microsc.* 2006;223:120-132.
26. Sadaphal P, Rao J, Comstock GW, et al. Image processing techniques for identifying *Mycobacterium tuberculosis* in Ziehl-Neelsen stains. *Int J Tuberc Lung Dis.* 2008;12:579-582.
27. Khutlang R, Krishnan S, Whitelaw A, et al. Automated detection of tuberculosis in Ziehl-Neelsen-stained sputum smears using two one-class classifiers. *J Microsc.* 2010;237:96-102.
28. Osibote OA, Dendere R, Krishnan S, et al. Automated focusing in bright-field microscopy for tuberculosis detection. *J Microsc.* 2010;240:155-163.
29. Lewis JJ, Chihota VN, van der Meulen M, et al. "Proof-of-concept" evaluation of an automated sputum smear microscopy system for tuberculosis diagnosis. *PLoS One.* 2012;7:e50173.
30. Chang J, Arbeláez P, Switz N, et al. Automated tuberculosis diagnosis using fluorescence images from a mobile microscope. *Med Image Comput Comput Assist Interv.* 2012;15:345-352.
31. Zingue D, Weber P, Soltani F, et al. Automatic microscopic detection of mycobacteria in sputum: a proof-of-concept. *Sci Rep.* 2018;8:11308.
32. Wang HS, Liang WY. Automatic *Mycobacterium tuberculosis* detection using simple image processing with artificial intelligence. *J Pathol Inform.* 2019;10:S40.
33. Russell MJ. *Autofocusing and Image Segmentation in Microscopy for Automatic Detection of Tuberculosis in Sputum Smears* [Master's thesis]. Cape Town, South Africa: University of Cape Town; 2006. <https://open.uct.ac.za/handle/11427/3198>. Accessed June 8, 2020.
34. Tavolara TE, Niazi MKK, Beamer G, et al. Segmentation of *Mycobacterium tuberculosis* bacilli clusters from acid-fast stained lung biopsies: a deep learning approach. *SPIE Proc.* 2020;113200E-1-113200E-7.
35. Wu RI, Mark EJ, Hunt JL. Staining for acid-fast bacilli in surgical pathology: practice patterns and variations. *Hum Pathol.* 2012;43:1845-1851.
36. Smith KP, Kirby JE. Image analysis and artificial intelligence in infectious disease diagnostics. *Clin Microbiol Infect.* 2020;26:1318-1323.
37. Rhoads DD. Computer vision and artificial intelligence are emerging diagnostic tools for the clinical microbiologist. *J Clin Microbiol.* 2020;58:e00511-20.
38. Peiffer-Smadja N, Dellièrre S, Rodriguez C, et al. Machine learning in the clinical microbiology laboratory: has the time come for routine practice? *Clin Microbiol Infect.* 2020;26:1300-1309.

OXFORD UNIVERSITY PRESS LICENSE TERMS AND CONDITIONS

Oct 22, 2021

This Agreement between Liron Pantanowitz ("You") and Oxford University Press ("Oxford University Press") consists of your license details and the terms and conditions provided by Oxford University Press and Copyright Clearance Center.

License Number 5174230672144

License date Oct 22, 2021

Licensed content
publisher Oxford University Press

Licensed content
publication American Journal of Clinical Pathology

Licensed content
title Artificial Intelligence–Based Screening for Mycobacteria in Whole-Slide Images of Tissue Samples

Licensed content
author Pantanowitz, Liron; Wu, Uno

Licensed content
date Feb 2, 2021

Type of Use Thesis/Dissertation

Institution name

Title of your work	DIGITAL PATHOLOGY & ARTIFICIAL INTELLIGENCE: FEASIBILITY FOR CLINICAL PRACTICE
Publisher of your work	University of Witwatersrand
Expected publication date	Jan 2022
Permissions cost	0.00 USD
Value added tax	0.00 USD
Total	0.00 USD
Title	DIGITAL PATHOLOGY & ARTIFICIAL INTELLIGENCE: FEASIBILITY FOR CLINICAL PRACTICE
Institution name	University of Witwatersrand
Expected presentation date	Jan 2022
Order reference number	Wits PhD
Requestor Location	Liron Pantanowitz NCRC Building 35 2800 Plymouth Rd Ann Arbor, MI 48109 United States Attn: Liron Pantanowitz
Publisher Tax ID	GB125506730

Total 0.00 USD

Terms and Conditions

STANDARD TERMS AND CONDITIONS FOR REPRODUCTION OF MATERIAL FROM AN OXFORD UNIVERSITY PRESS JOURNAL

1. Use of the material is restricted to the type of use specified in your order details.
2. This permission covers the use of the material in the English language in the following territory: world. If you have requested additional permission to translate this material, the terms and conditions of this reuse will be set out in clause 12.
3. This permission is limited to the particular use authorized in (1) above and does not allow you to sanction its use elsewhere in any other format other than specified above, nor does it apply to quotations, images, artistic works etc that have been reproduced from other sources which may be part of the material to be used.
4. No alteration, omission or addition is made to the material without our written consent. Permission must be re-cleared with Oxford University Press if/when you decide to reprint.
5. The following credit line appears wherever the material is used: author, title, journal, year, volume, issue number, pagination, by permission of Oxford University Press or the sponsoring society if the journal is a society journal. Where a journal is being published on behalf of a learned society, the details of that society must be included in the credit line.
6. For the reproduction of a full article from an Oxford University Press journal for whatever purpose, the corresponding author of the material concerned should be informed of the proposed use. Contact details for the corresponding authors of all Oxford University Press journal contact can be found alongside either the abstract or full text of the article concerned, accessible from www.oxfordjournals.org Should there be a problem clearing these rights, please contact journals.permissions@oup.com
7. If the credit line or acknowledgement in our publication indicates that any of the figures, images or photos was reproduced, drawn or modified from an earlier source it will be necessary for you to clear this permission with the original publisher as well. If this permission has not been obtained, please note that this material cannot be included in your publication/photocopies.
8. While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by Oxford University Press or by Copyright Clearance Center (CCC)) as provided in CCC's Billing and Payment terms and conditions. If

full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and Oxford University Press reserves the right to take any and all action to protect its copyright in the materials.

9. This license is personal to you and may not be sublicensed, assigned or transferred by you to any other person without Oxford University Press's written permission.

10. Oxford University Press reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

11. You hereby indemnify and agree to hold harmless Oxford University Press and CCC, and their respective officers, directors, employs and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

12. Other Terms and Conditions:

v1.4

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

CHAPTER 3: PUBLISHED ARTICLE 2

Pantanowitz L, Hartman D, Qi Y, Cho EY, Suh B, Paeng K, Dhir R, Michelow P, Hazelhurst S, Song SY, Cho SY. Accuracy and efficiency of an artificial intelligence tool when counting breast mitoses. *Diagnostic Pathology*, 2020; 51(1):80.

<https://pubmed.ncbi.nlm.nih.gov/32622359/>

Abstract

Background: The mitotic count in breast carcinoma is an important prognostic marker. Unfortunately substantial inter- and intra-laboratory variation exists when pathologists manually count mitotic figures. Artificial intelligence (AI) coupled with whole slide imaging offers a potential solution to this problem. The aim of this study was to accordingly critique an AI tool developed to quantify mitotic figures in whole slide images of invasive breast ductal carcinoma.

Methods: A representative H&E slide from 320 breast invasive ductal carcinoma cases was scanned at 40x magnification. Ten expert pathologists from two academic medical centers labeled mitotic figures in whole slide images to train and validate an AI algorithm to detect and count mitoses. Thereafter, 24 readers of varying expertise were asked to count mitotic figures with and without AI support in 140 high-power fields derived from a separate dataset. Their accuracy and efficiency of performing these tasks were calculated and statistical comparisons performed.

Results: For each experience level the accuracy, precision and sensitivity of counting mitoses by users improved with AI support. There were 21 readers (87.5%) that identified more mitoses using AI support and 13 reviewers (54.2%) that decreased the quantity of falsely flagged mitoses with AI. More time was spent on this task for most participants when not provided with AI support. AI assistance resulted in an overall time savings of 27.8%.

Conclusions: This study demonstrates that pathology end-users were more accurate and efficient at quantifying mitotic figures in digital images of invasive breast carcinoma with the aid of AI. Higher inter-pathologist agreement with AI assistance suggests that such algorithms can also help standardize practice. Not surprisingly, there is much enthusiasm in pathology regarding the prospect of using AI in routine practice to perform mundane tasks such as counting mitoses.

RESEARCH

Open Access



Accuracy and efficiency of an artificial intelligence tool when counting breast mitoses

Liron Pantanowitz^{1,2*}, Douglas Hartman¹, Yan Qi³, Eun Yoon Cho⁴, Beomseok Suh⁵, Kyunghyun Paeng⁵, Rajiv Dhir¹, Pamela Michelow², Scott Hazelhurst⁶, Sang Yong Song^{4†} and Soo Youn Cho^{4†}

Abstract

Background: The mitotic count in breast carcinoma is an important prognostic marker. Unfortunately substantial inter- and intra-laboratory variation exists when pathologists manually count mitotic figures. Artificial intelligence (AI) coupled with whole slide imaging offers a potential solution to this problem. The aim of this study was to accordingly critique an AI tool developed to quantify mitotic figures in whole slide images of invasive breast ductal carcinoma.

Methods: A representative H&E slide from 320 breast invasive ductal carcinoma cases was scanned at 40x magnification. Ten expert pathologists from two academic medical centers labeled mitotic figures in whole slide images to train and validate an AI algorithm to detect and count mitoses. Thereafter, 24 readers of varying expertise were asked to count mitotic figures with and without AI support in 140 high-power fields derived from a separate dataset. Their accuracy and efficiency of performing these tasks were calculated and statistical comparisons performed.

Results: For each experience level the accuracy, precision and sensitivity of counting mitoses by users improved with AI support. There were 21 readers (87.5%) that identified more mitoses using AI support and 13 reviewers (54.2%) that decreased the quantity of falsely flagged mitoses with AI. More time was spent on this task for most participants when not provided with AI support. AI assistance resulted in an overall time savings of 27.8%.

Conclusions: This study demonstrates that pathology end-users were more accurate and efficient at quantifying mitotic figures in digital images of invasive breast carcinoma with the aid of AI. Higher inter-pathologist agreement with AI assistance suggests that such algorithms can also help standardize practice. Not surprisingly, there is much enthusiasm in pathology regarding the prospect of using AI in routine practice to perform mundane tasks such as counting mitoses.

Keywords: Artificial intelligence, Breast, Carcinoma, Counting, Tumor grade, Digital pathology, Informatics, Mitosis, Whole slide imaging

* Correspondence: lpantanowitz@gmail.com

†Sang Yong Song and Soo Youn Cho share senior authorship on this paper.

¹Department of Pathology, University of Pittsburgh Medical Center Cancer Pavilion, Suite 201, 5150 Centre Ave, Pittsburgh, PA 15232, USA

²Department of Anatomical Pathology, University of the Witwatersrand and National Health Laboratory Services, Johannesburg, South Africa

Full list of author information is available at the end of the article



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Background

Handling breast cancer specimens is common in pathology practice. Rendering a pathology report after processing these specimens not only requires an accurate diagnosis, but in the case of invasive carcinoma also requires pathologists to assign the correct histologic tumor grade. A key component of the Nottingham (or modified Scarff-Bloom-Richardson) grading system for invasive breast carcinoma includes the mitotic count [1]. A mitotic count per 10 high-power fields (HPFs) of 0–7 is scored 1, 8–15 is scored 2, and greater than or equal to 16 is given a score of 3. This proliferation activity in breast carcinoma is an important prognostic marker [2]. Some studies have shown that the mitotic count is even a better marker than Ki67 (proliferation index) at selecting patients for certain therapy such as tamoxifen [3].

Counting mitotic figures in hematoxylin and eosin (H&E) stained histology sections is a task typically performed by pathologists while they visually examine a glass slide using a conventional light microscope. Unfortunately, there is substantial inter- and intra-laboratory variation with manual grading of breast cancer in routine pathology practice [4]. This is not surprising, as manually counting mitotic figures by pathologists is subjective and suffers from low reproducibility. Manually counting mitoses can take a pathologist around 5–10 min to perform [5]. Sometimes it may be difficult to discern a mitotic figure from a cell undergoing degeneration, apoptosis or necrosis. There are also differences of opinion on how best to count mitotic figures [6, 7]. The reason for this controversy is that the mitotic activity index depends on the number of mitoses counted in a predefined area (usually in mm²) or within a certain number of HPFs that may vary depending on a microscope's lenses and widefield microscopy view.

Artificial intelligence (AI) coupled with whole slide imaging offers a potential solution to the aforementioned problem. If developed and deployed successfully, an AI-based tool could potentially automate the task of counting mitotic figures in breast carcinoma with better accuracy and efficiency. To date, investigators have validated that making a histopathologic diagnosis in breast specimens can be reliably performed on a whole slide image (WSI) [8]. Moreover, using WSIs to manually count mitoses in breast cancer is reported to be reliable and reproducible [9, 10]. Hanna et al. showed that counting mitotic figures in WSIs outperformed counts using glass slides, albeit this took readers longer using WSI [11]. Several studies have been published showing that digital image analysis can successfully automate the quantification of mitoses [12–18].

Clearly, there is great potential for leveraging digital pathology and AI [19]. AI can benefit pathologists practicing in high, middle and low income countries,

especially with the rise in cancer and shortage of anatomical pathologists [20]. However, AI applications in healthcare have not been vigorously validated for reproducibility, generalizability and in the clinical setting [21]. Moreover, hardly any pathology laboratories are currently using AI tools on a routine basis. To the best of our knowledge, there have been no studies addressing whether an AI-based algorithm actually improves pathologist accuracy and efficiency when scoring mitotic figures. The aim of this study was to accordingly critique an AI tool developed to detect and quantify mitotic figures in breast carcinoma.

Methods

Figure 1 depicts a flow chart of the methodology and datasets employed in developing and validating the AI-based tool utilized in this study to quantify mitotic figures in digital images of invasive breast carcinoma.

Datasets

A total of 320 invasive breast ductal carcinoma cases with an equal distribution of grades were selected. Half of these cases were from the archives of the University of Pittsburgh Medical Center (UPMC) in the USA and the rest obtained from Samsung Medical Center (SMC) in Seoul, South Korea. Nearly all of the cases were from females (1 case was from a male with breast cancer). The average patient age was 54.7 years. All cases included were mastectomies with the following range of tumor stages: stage IA (23.6%), IB (7.1%), IIA (31.4%), IIB (23.6%), IIC (0.7%), IIIA (6.4%), IV (0.7%), and data unavailable in 9 cases (6.4%). Table 1 provides a summary of the cancer grade, hormone receptor and HER2 status for enrolled cases (with available data). The average Ki-67 index was 38.3% (Mdn = 34.5%, range 3.0–99.0%). This result was only available in 80 cases, and this subset of cases had higher mitosis scores ($n = 23$ score 2, $n = 48$ score 3) and Nottingham grades ($n = 34$ grade 2, $n = 41$ grade 3). The average proliferation index was accordingly skewed in this subset and higher than would be expected for a typical mixed breast cancer population [22].

A representative H&E glass slide from each case was scanned. At UPMC slides were scanned at 40x magnification (0.25 $\mu\text{m}/\text{pixel}$ resolution) using an Aperio AT2 scanner (Leica Biosystems Inc., Buffalo Grove, IL, USA). At SMC slides were digitized at 40x magnification (0.2 $\mu\text{m}/\text{pixel}$ resolution) using a 3D Histech P250 instrument (3DHISTECH, Budapest, Hungary). All acquired whole slide image (WSI) files were de-identified. The AI training dataset was comprised of 60 WSIs from UPMC and 60 WSIs from SMC, which provided 16,800 grids (1 grid = $\frac{1}{4}$ high-power field [HPF]). One HPF is equivalent to 0.19 mm². The AI validation dataset,

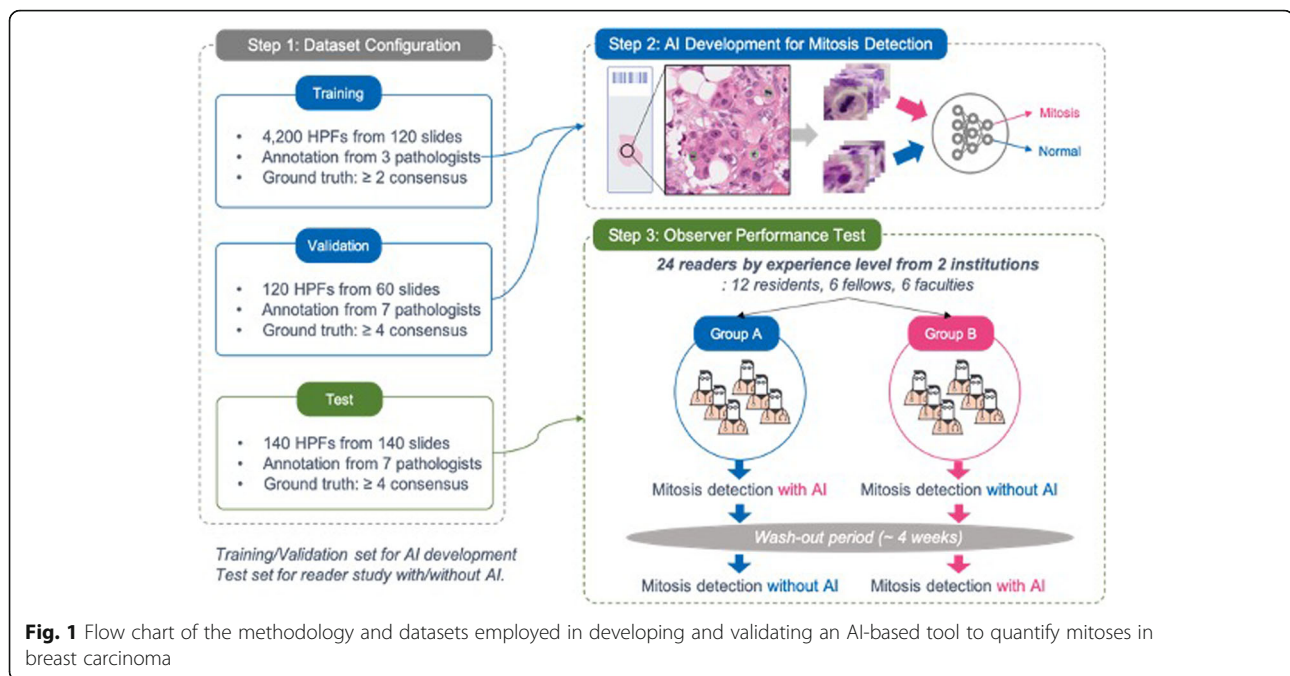


Fig. 1 Flow chart of the methodology and datasets employed in developing and validating an AI-based tool to quantify mitoses in breast carcinoma

Table 1 Profile of invasive ductal carcinoma cases enrolled in the study

Reported breast carcinoma parameters		%
Mitosis Score	1	21.4%
	2	31.4%
	3	47.1%
Nottingham Grade	1	7.9%
	2	46.4%
	3	45.7%
ER	Not available	5.0%
	Negative	25.7%
	Positive	69.3%
PR	Not available	5.0%
	Negative	32.1%
	Positive	62.9%
HER2/neu (IHC status)	Not available	5.0%
	Negative	59.3%
	Equivocal	9.3%
	Weakly positive	1.4%
	Positive	25.0%
HER2/neu (FISH status)	Not available	89.3%
	Negative	10.0%
	Positive	0.7%

ER estrogen receptor, FISH fluorescence in situ hybridization, HER2 human epidermal growth factor receptor 2, IHC immunohistochemistry, PR progesterone receptor

comprised of another 30 WSIs from UPMC and 30 WSIs from SMC, was used to generate 120 HPFs for annotation. A separate dataset (70 WSIs from UPMC and 70 WSIs from SMC) was subsequently used for a reader study where each WSI file was randomly broken up into 140 representative digital patches (HPFs). Users interacted with individual patches on a computer monitor. The dataset used for analytical validation of the algorithm was different from the dataset selected for the clinical validation study.

Training (deep learning algorithm)

A deep learning algorithm (Lunit Inc., Seoul, South Korea) was employed for the automated detection of mitoses in digital images [23]. The AI algorithm was trained on an independent dataset, that consisted of 16,800 digital image patches from 120 WSIs (half from UPMC and half from SMC). Three expert pathologists annotated mitoses to construct the ground truth for training. The mitotic figures, which were the consensus of at least two of these pathologists, were used to train the AI algorithm. The algorithm was based on Faster RCNN [24] by ResNet-101 [25] backbone network that has pre-trained weights. The down sampling ratio was 8 and feature maps from the first stage were cropped and resized at 14×14 and then max pooled to 7×7 for the second stage classifier. Anchor size was 128×128 with a single fixed ratio. The number of proposals at the first stage was 2000 to enable a very dense sampling of proposal boxes. Then, box IOU based NMS was performed for post-processing. Various input data augmentation methods such as contrast, brightness, jittering, flip and

rotation were performed to build a robust AI algorithm. To select the final model for our reader study, the deep learning algorithm was validated on a separate dataset. Employing the validation dataset we achieved 0.803 mean AP (mAP) which demonstrates good performance. The mAP represents the area under the precision recall curve. A precision recall curve was used to calculate the mAP instead of AUC, because of the large class imbalance (i.e., many non-mitotic cells).

Ground truth

Seven expert pathologists (4 from UPMC and 3 from SMC) annotated (labeled) mitotic figures in 140 digital image patches using a web-based annotation tool. The tool displayed image patches of breast carcinoma at high magnification, in which clicking on cells automatically generated a square box that annotated the specified cell (i.e. with the mitotic figure present). It required around 10 s to annotate mitotic figures per patch. Pathologist consensus was used to establish ground truth, where agreement of at least 4/7 pathologists was required for each image. Whilst there is no published data available to support the exact number of pathologists required to be in agreement to reach consensus, a consensus of 4 out of 7 was chosen for this study in order to utilize the highest number of cases ($n = 93$, 66.4%) while maintaining consensus among the majority of ground truth makers (57.1%). Table S1 shows the number of cases for each consensus level. Further, for 100% agreement the mitotic figures would likely be very obvious and thus too easy to detect, which would not be suitable to measure performance. Since prior studies have proven that WSI can be used for mitotic cell detection and offers similar reproducibility to the microscope [10, 26], we opted to use WSI and not glass slides for establishing the ground

truth in this study. Pathologists who annotated slides for ground truth generation did not participate in the subsequent reader study.

Observer performance test (OPT)

For the OPT (reader study), the accuracy and efficiency of mitotic cell detection was compared based on mitotic figure scores provided by humans and the AI algorithm. There were 12 readers at each institution (total of 24 reviewers) that varied in expertise/years of experience ($n = 6$ 2nd-4th year pathology residents/registrar, $n = 3$ fellows/post-residency trainees, and $n = 3$ board-certified pathologists). Table S2 summarizes the experience level of all participants involved in the study. Digital slides were presented to test takers in the form of 140 HPFs. Each HPF was equivalent to four digital image patches. There were two reader groups. In group 1 (no AI), readers were first shown HPFs and asked to manually select mitotic figures without AI support. In group 2 (with AI), readers were first shown HPFs where mitotic figures were pre-marked by the AI tool (Figure 2) and asked to accept/reject the algorithm's selection. Each group repeated this task, but now with/without AI employing a cross-over design to minimize sequential confounding bias. A washout period of 4 weeks was used to control for recall bias between re-reviews of each image. A web-based tool recorded user clicks on images and their time (in seconds) to perform this task. The OPT was replicated at UPMC and SMC institutions. All readers were trained prior to the start of the study, anonymized, and provided informed consent to participate. The readers were not formally asked to provide feedback about their user experience.

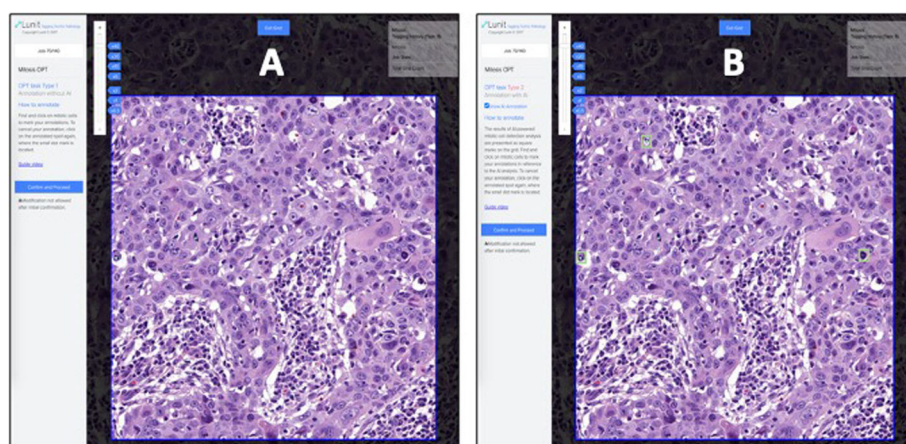


Fig. 2 Web-based tool showing a HPF of breast carcinoma. **a** Screenshot of the web-based tool used for the observer performance test without AI. The small green dots indicate mitotic figures marked by the reader. **b** Screenshot of the web-based tool used for the observer performance test with AI. The green boxes indicate mitotic figures detected by AI

Statistical Analysis

Accuracy of mitotic cell detection was calculated by comparing cells identified by reviewers to cells identified by the ground truth (i.e. consensus of at least 4 of the 7 ground truth makers). Accuracy was compared for reviews with and without AI support for each reviewer. The hypothesis being tested was that reviewer accuracy improves with AI support. To test this hypothesis a Pearson chi-square analysis was performed. For the OPT part of this study, true positive (TP), false positive (FP) and false negative (FN) were calculated with and without AI support. Precision for pathologists was calculated as $TP / (TP + FP)$. Sensitivity was calculated as $TP / (TP + FN)$. As true negatives (TN) represented not only cells, but also all of the white space where no cells were present in an image, TN greatly outnumber the combination of TP + FP + FN and therefore f-scores were calculated ($f\text{-score} = 2 * ((\text{sensitivity} * \text{precision}) / (\text{sensitivity} + \text{precision}))$). F-scores closer to 1 indicate perfect detection and precision. Since TN were not calculated, specificity was not possible to calculate.

Efficiency was calculated as seconds spent reviewing each case. The normality of the distribution of the time variable was examined using the Shapiro–Wilk normality test. As the data were not normally distributed, non-parametric statistical tests were used. Wilcoxon signed-rank test was used to compare time spent on the task of counting mitoses with and without AI support. We assumed that image reviews lasting longer than 10 min were outliers (e.g. indicative of an interruption) and thus

excluded. Out of the 6720 values in the dataset, 73 (1.1%) were accordingly excluded from analysis. Statistical comparisons were performed for time spent per case with and without AI support for each individual, for each user's experience level, and overall.

Statistical significance was assumed at $p < .05$. Analysis was performed using IBM SPSS Statistics 22 and Microsoft Excel 365.

Results

Accuracy and precision findings

A precision recall (PR) curve shows the algorithm's performance (Figure 3). This PR curve shows the relationship between positive predictive value and sensitivity for every possible cut-off. Akin to the area under a ROC curve (i.e. AUC), the area under the PR curve is large indicating the high recall and precision value of the algorithm at specific cut-offs. Figure 4 shows the accuracy and precision of mitotic cell detection with and without the use of AI support. For each experience level the accuracy and precision were higher with AI support. Table 2 with Chi-square results confirmed that accurate mitotic cell detection was significantly higher with the use of AI support for each experience level. Table S3 shows the individual reviewer accuracy results. Of note, all but one reviewer had higher accuracy with the support of AI. Of the 23 reviewers with improved accuracy, 20 (87%) had a statistically significant increase. Table 3 demonstrates TP, FP and FN values for readers (Table S4 shows individual reviewer results). There were 21 out of the 24 readers

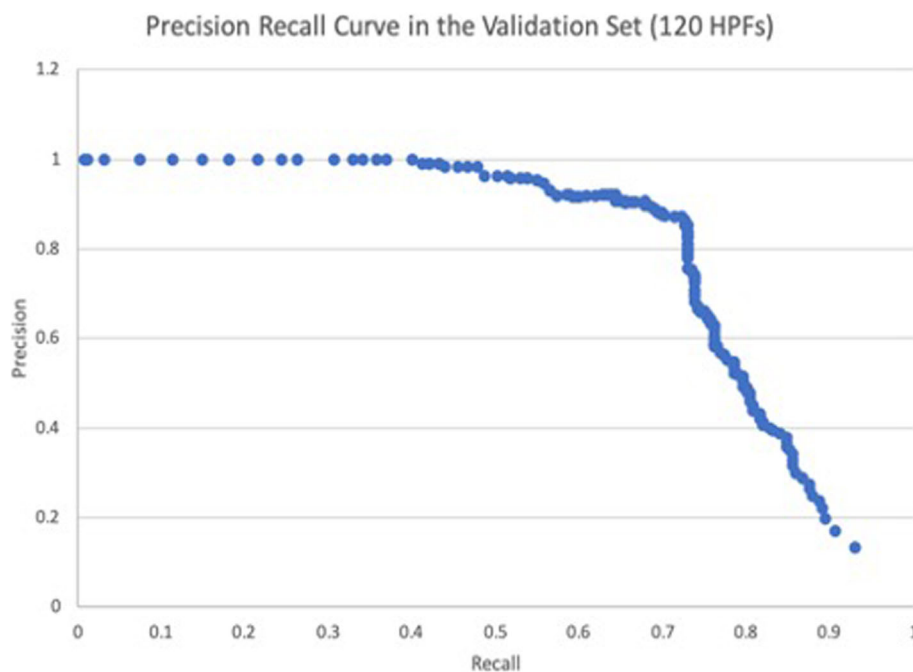


Fig. 3 Algorithm performance for mitotic figure detection in the analytical validation dataset

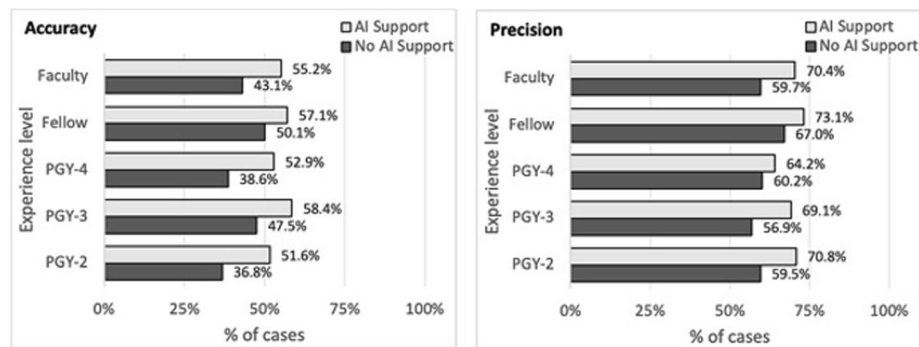


Fig. 4 Accuracy and precision with and without AI support per user experience level

(87.5%) that identified more mitoses using AI support. Further, 13 reviewers (54.2%) decreased the quantity of falsely flagged mitoses (FP) using AI support, and 21 (87.5%) decreased the quantity of mitoses that were missed (FN) using AI support. There were six reviewers that falsely detected 100 or more additional mitoses (FP) when screening cases without AI support. Table 3 shows that the number of FPs detected with the use of AI support (2899) is lower than without the use of AI support (3587).

Sensitivity for mitotic cell detection increased with the use of AI support for each experience level (Table S5). Sensitivity for mitotic cell detection per individual reviewer was higher for all but 3 reviewers. Precision for mitotic cell detection also increased with the use of AI support for each experience level (Table S6). Sixteen of the 24 reviewers (66.7%) had increased precision with AI support. The f-score (Table S7) for mitotic cell detection without the use of AI support was 0.61, and with the use of AI support was 0.71. The higher f-score with the use of AI suggests that AI support improves overall precision and TP detection of mitotic cells. Cases with AI support also had higher f-scores for each experience level, with 23 of the 24 reviewers (95.8%) demonstrating a higher f-score with AI support. The datasets utilized included only the overall grade (i.e. sum of percent tubules, nuclear pleomorphism and mitoses/10 HPF) for all breast cancers and no details of the exact mitotic figures (i.e. score 1, 2 or 3) for each case. Therefore, we were unable to investigate

whether any change in the number of mitoses scored in this study may have altered the grade.

Efficiency findings

A Wilcoxon signed-rank test indicated that more time was spent on detecting mitotic cells without the use of AI support (median = 36.00 s) than with AI support (median = 26.00 s), $Z = -14.759$, $p < .001$, $r = .25$. Overall, this represents a time savings of 27.8%. Irrespective of whether readers started counting mitoses with or without AI support, nearly all of them read faster with AI assistance, but this was not statistically different. Figure 5 shows the median time spent detecting mitoses with and without AI support by reader experience level. Despite experience level, most participants spent less time detecting mitotic cells with the use of AI support. Fellows had the largest decline, with a median of 44 s spent without the aid of AI compared to 16 s with AI support. The only experience level that had a longer median time spent with AI support was postgraduate year (PGY)-4 users. Table 4 summarizes the median time spent and statistical results per user's experience level with and without AI support (Table S8 shows individual reviewer results).

Conclusions

There are formidable challenges with successfully translating AI in healthcare [10, 19, 26]. Some of these

Table 2 Accuracy by experience level

User Experience Level	No AI Support	With AI Support	Improved Accuracy with AI support?	χ^2 (degrees of freedom)	<i>p</i> -value
PGY-2 (<i>n</i> = 4)	36.8%	51.6%	Yes	89.30 (1)	<.001
PGY-3 (<i>n</i> = 4)	47.5%	58.4%	Yes	53.12 (1)	<.001
PGY-4 (<i>n</i> = 4)	38.6%	52.9%	Yes	87.13 (1)	<.001
Fellow (<i>n</i> = 6)	50.1%	57.1%	Yes	29.82 (1)	<.001
Faculty (<i>n</i> = 6)	43.1%	55.2%	Yes	89.84 (1)	<.001
Overall	43.9%	55.2%	Yes	320.61 (1)	<.001

PGY postgraduate year

Table 3 True positive (TP), false positive (FP), and false negative (FN) values for mitotic cell detection

User Experience Level	No AI support			With AI support		
	TP	FP	FN	TP	FP	FN
PGY-2 (n = 4)	749	509	779	1003	414	525
PGY-3 (n = 4)	1135	861	393	1208	539	320
PGY-4 (n = 4)	793	525	735	1149	642	379
Fellow (n = 6)	1524	751	768	1659	611	633
Faculty (n = 6)	1395	941	897	1647	693	645
Overall	5596	3587	3572	6666	2899	2502

PGY postgraduate year

challenges include technical difficulties, complex implementations, data ownership issues, lack of reimbursement, delayed regulatory approval, ethical concerns, and overcoming human trepidation regarding AI (e.g. mistrust related to the 'black box' phenomenon of AI). Bairnov et al. showed that an AI-based decision support tool in Radiology had significant differences with accuracy and inter-operator variability depending on how AI was deployed (i.e. sequential or independent workflow) [21]. To the best of our knowledge, no studies have been published examining the interaction of pathology end users with AI to determine the pros and cons of using AI to assist with counting mitoses. Such studies would provide much needed translational evidence that could help develop recommendations and guidelines for the safe and effective use of AI in routine diagnostic Anatomical Pathology workflow.

This cross validation study demonstrates that pathology end-users were more accurate and efficient at quantifying mitotic figures in digital images of invasive breast carcinoma with the aid of an AI tool that detects mitoses. These data show that the accuracy, sensitivity, precision, and f-scores all increased for each participant experience level with the use of AI support. Readers in both groups had higher inter-pathologist agreement with AI assistance, suggesting that AI can help standardize practice and perhaps result in more reproducible

diagnoses. Very few participants unexpectedly had a lower accuracy performance with AI support. The results of this study showed that only 54.2% of reviewers decreased the quantity of falsely flagged mitoses using AI support. The reason why false positives were not reduced across all readers with AI support could be that they missed annotated mitotic figures because they were not clearly visible in the user interface or that some readers may not have believed the AI results. A detailed analysis of the sessions from these individuals showed that for some cases they spent an unusually long time counting mitoses (e.g. 451 s in one case with AI support, but only 15 s on the same case without AI support). This likely points to distraction more than AI causing an actual delay and it is uncertain if these outliers skewed the data. With regard to improved efficiency, the use of AI resulted in a 27.8% decrease in time for mitotic cell detection. In other words, for every 1 h spent searching for cells with mitotic figures without AI support, roughly 16.7 min could be saved using AI support. Nearly every subgroup of participants had faster reading speeds with the use of AI (PGY-4 was the exception). Overall, 66.7% of pathologists read faster with AI (statistically significantly faster for 33.3%). For pathology trainees, use of AI support resulted in faster reads for 83.3% of residents/registrar (statistically significantly faster for 25.0%) and 83.3% of fellows (all 83.3% statistically significantly faster).

Methods to automatically detect mitoses in breast cancer images were introduced in the literature several decades ago [27]. Despite limited access to large digital datasets and prior to the availability of today's computer processing power, many early image analysis projects demonstrated the feasibility of using computers to assist in counting mitoses [28, 29]. Although some of these first generation algorithms provided promising results, they were not yet suitable for clinical practice. Since then, with the advent of newer technologies including WSI, deep learning methods, graphics processing units and cloud computing we have witnessed a new generation of AI-based algorithms that are able to automate mitosis detection with impressive performance [16, 30–36]. Several international challenges using public datasets catalyzed the development of these sophisticated AI tools [37, 38], including algorithms to predict breast tumor proliferation [39]. The Lunit algorithm utilized in this study to automate mitosis counting in breast carcinoma WSIs integrates three modules: (i) image processing to handle digital slides (e.g. tissue region and patch extraction, region of interest detection, stain normalization), (ii) deep learning mitosis detection network (based on Residual Network or ResNet architecture), and (iii) a proliferation score prediction module [23]. For the Tumor Proliferation Assessment Challenge in 2016 (TUPAC16; <http://tupac.tue-image.nl/>), Lunit

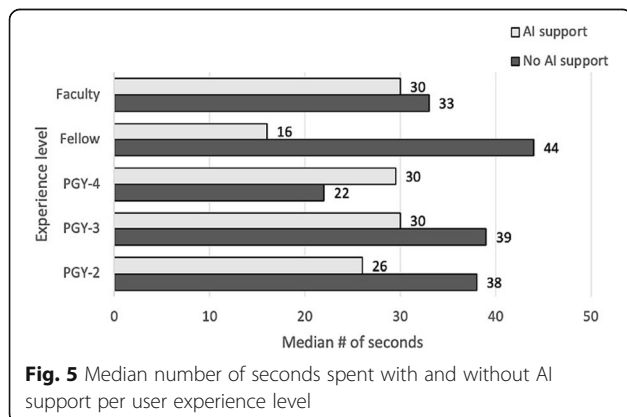


Table 4 Median time to count mitoses by study participant experience level

User Experience Level	Median # of seconds		AI or no AI faster?	Z	p-value	r
	No AI support	With AI support				
PGY-2 (n = 4)	38.00	26.00	AI	-8.799	<.001	.37
PGY-3 (n = 4)	39.00	30.00	AI	-3.290	.001	.14
PGY-4 (n = 4)	22.00	29.50	No AI	-3.058	.002	.13
Fellow (n = 6)	44.00	16.00	AI	-16.730	<.001	.58
Faculty (n = 6)	33.00	30.00	AI	-2.584	.010	.09
Overall	36.00	26.00	AI	-14.759	<.001	.25

r effect size, PGY postgraduate year

won all tasks including the prediction of mitosis grading. For this specific task their method achieved a Cohen's kappa score of $\kappa = 0.567$, 95% CI [0.464, 0.671] between the predicted scores and the ground truth [17].

In general, mitotic figures are detectable in H&E stained tissue sections due to their hyperchromatic appearance and characteristic shapes. However, it is plausible that mitoses may be missed by humans and/or even AI algorithms due to tissue or imaging artifacts. To address this, using a biomarker such as Phosphorylated Histone H3 (PHH3) may have helped objectively confirm mitotic figures [40, 41]. Even though overall accuracy for readers in the OPT study was determined to be 55.2%, with AI support this was still more sensitive than counting mitotic figures manually. Further, contrary to classifying mitoses into scores 1, 2, and 3 for actual diagnostic purposes, this study was aimed at finding individual mitotic cells in a simulated format, which is expected to have relatively lower performance that could have caused missed or incorrect mitotic figure detection. Davidson et al. have shown that while pathologists' reproducibility is similar for Nottingham grade using glass slides or WSI, there is still slightly lower intraobserver agreement because grading breast cancer using digital WSI is more challenging [42]. Another limitation of our study was not standardizing the monitors used for annotation and the reader study. However, Norgan et al. showed that manual mitotic figure enumeration by pathologists was not affected by medical-grade versus commercial off-the-shelf displays [43]. In this study we did not equate a glass slide HPF with a digital HPF. Indeed, currently the HPF is typically used in manual microscopy with glass slides when quantifying mitoses (e.g. breast mitoses are evaluated using 10 HPFs at 400x magnification) [44]. However, this HPF at 400x on a glass slide is unlikely to be equivalent to a digital HPF at "40x view" view in a WSI [45].

As verified by this study, expected benefits of adopting AI in pathology practice include automation, elimination of tedious tasks, improved accuracy, and efficiency. Not surprisingly, there is much enthusiasm in pathology regarding the prospect of using AI in routine practice.

Interestingly, some of the trainees involved in this study expressed their gratitude for being invited to participate because of the opportunity to experience working with AI first hand. Of course, there is much to still be learned before successfully embedding AI into routine workflows. If AI is indeed more accurate than humans at counting mitoses we will need to determine how this impacts patient outcomes and whether man-made scoring systems may need to be revised.

Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s13000-020-00995-z>.

Additional file 1: Table S1. Number of cases based on consensus among ground truth makers. **Table S2.** Experience level of participants involved in the OPT component of the study. **Table S3.** Individual accuracy reviewer results for the OPT. **Table S4.** Individual reviewer TP, FP, FN mitotic cell detection results for the OPT. **Table S5.** Sensitivity results by experience level and individual reviewer for the OPT. **Table S6.** Precision results by experience level and individual reviewer for the OPT. **Table S7.** F-scores by experience level and individual reviewer for the OPT. **Table S8.** Individual reviewer results for time spent during the OPT.

Abbreviations

AI: Artificial intelligence; AUC: Area under ROC curve; FN: False negative; FP: False positive; H&E: Hematoxylin and eosin; HER2: Receptor tyrosine-protein kinase erbB-2; HPF: High-power field; IBM: International Business Machines Corporation; IOU: Intersection Over Union; mAP: Mean AP (area under precision recall curve); Mdn: Median; NMS: Non-maximum suppression; OPT: Observer performance test; PGY: Postgraduate year; PHH3: Phosphorylated Histone H3; PR: Precision recall; RCNN: Regions with convolutional neural networks; SMC: Samsung Medical Center; TP: True positive; TUPAC16: Tumor Proliferation Assessment Challenge in 2016; UPMC: University of Pittsburgh Medical Center; USA: United States of America; WSI: Whole slide image

Acknowledgements

We thank all of the participants in this study. We also thank Colleen Vrbn from Analytical Insights, LLC for her help with our statistical analysis.

Authors' contributions

LP – study conception, methodology, project administration, study coordination, informatics support, literature review, annotation; data collection, data curation, data analysis, manuscript preparation; YQ – literature search, data analysis, data interpretation, manuscript writing; SYC – study design, study coordination, data collection, data curation, expert review, manuscript review; EYC – study design, data collection, data curation, expert review; SYS – study conception, study coordination, expert review, manuscript review; BS – study conception, study coordination, study design; KP – study methodology, study coordination, data analysis, manuscript

writing; RD - methodology, study coordination, annotation; data collection, manuscript preparation; PM - study design, manuscript preparation; SH - study design, manuscript preparation; DH - annotation; data collection, manuscript preparation. The authors read and approved the final manuscript.

Funding

Lunit funded this study via a sponsored research agreement which was used for slide and data procurement, scanning and data generation.

Availability of data and materials

The datasets generated and/or analyzed during this study are not publicly available because they are saved on private servers, but may be available from the corresponding author on reasonable request.

Ethics approval and consent to participate

Institutional Review Board approval was obtained for this study (University of Pittsburgh, PA, USA PRO18010404; University of Witwatersrand, Johannesburg, South Africa clearance certificate M191003). All readers for the OPT provided informed consent to participate.

Competing interests

Liron Pantanowitz is a consultant for Hamamatsu and serves on the medical advisory board for Ibex. Beomseok Suh and Kyunghyun Paeng work for Lunit. Every effort was made to avoid a conflict of interest that could potentially influence the study conclusions.

Author details

¹Department of Pathology, University of Pittsburgh Medical Center Cancer Pavilion, Suite 201, 5150 Centre Ave, Pittsburgh, PA 15232, USA. ²Department of Anatomical Pathology, University of the Witwatersrand and National Health Laboratory Services, Johannesburg, South Africa. ³School of Medicine, University of Pittsburgh, Pittsburgh, PA, USA. ⁴Department of Pathology, Samsung Medical Center, Seoul, South Korea. ⁵Lunit, Seoul, South Korea. ⁶School of Electrical & Information Engineering and Sydney Brenner Institute for Molecular Bioscience, University of the Witwatersrand, Johannesburg, South Africa.

Received: 24 May 2020 Accepted: 25 June 2020

Published online: 04 July 2020

References

- Elston CW, Ellis IO. Pathological prognostic factors in breast cancer. I. the value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology*. 1991;19(5):403–10.
- Chang JM, McCullough AE, Dueck AC. Back to basics: traditional Nottingham grade mitotic counts alone are significant in predicting survival in invasive breast carcinoma. *Ann Surg Oncol*. 2015;22(Suppl. 3):S509–15.
- Beelen K, Opdam M, Severson T, Koornstra R, Vincent A, Wesseling J, Sanders J, Vermorken J, van Diest P, Linn S. Mitotic count can predict tamoxifen benefit in postmenopausal breast cancer patients while Ki67 score cannot. *BMC Cancer*. 2018;18(1):761.
- van Doijewert C, van Diest PJ, Willems SM, Kuijpers CCHJ, van der Wall E, Overbeek LH, Deckers IAG. Significant inter- and intra-laboratory variation in grading of invasive breast cancer: a nationwide study of 33,043 patients in the Netherlands. *Int J Cancer*. 2020;146(3):769–80.
- Veta M. Breast Cancer Histopathology Image Analysis. The Netherlands: PhD Thesis, Utrecht University; 2014. Chapter 5:61–88.
- Facchetti F. A proposal for the adoption of a uniform metrical system for mitosis counting. *Int J Surg Pathol*. 2005;13(2):157–9.
- Yigit N, Gunal A, Kucukodaci Z, Karslioglu Y, Onguru O, Ozcan A. Are we counting mitoses correctly? *Ann Diagn Pathol*. 2013;17(6):536–9.
- Al-Janabi S, Huisman A, Willems SM, Van Diest PJ. Digital slide images for primary diagnostics in breast pathology: a feasibility study. *Hum Pathol*. 2012;43(12):2318–25.
- Al-Janabi S, van Slooten HJ, Visser M, van der Ploeg T, van Diest PJ, Jiwa M. Evaluation of mitotic activity index in breast cancer using whole slide digital images. *PLoS One*. 2013;8(12):e82576.
- Wei BR, Halsey CH, Hoover SB, Puri M, Yang HH, Gallas BD, et al. Agreement in histological assessment of mitotic activity between microscopy and digital whole slide images informs conversion for clinical diagnosis. *Acad Pathol*. 2019;6:2374289519859841.
- Hanna M, Xing J, Monaco SE, Hartman D, Pantanowitz L. Evaluation of diagnostic concordance between manual mitotic figure counting on glass slides versus whole slide images. *J Pathol Inform*. 2017;8:26.
- Malon C, Brachtel E, Cosatto E, Graf HP, Kurata A, Kuroda M, Meyer JS, Saito A, Wu S, Yagi Y. Mitotic figure recognition: agreement among pathologists and computerized detector. *Anal Cell Pathol (Amst)*. 2012;35(2):97–100.
- Cireşan DC, Giusti A, Gambardella LM, Schmidhuber J. Mitosis detection in breast cancer histology images with deep neural networks. *Med Image Comput Comput Assist Interv*. 2013;16(Pt 2):411–8.
- Roux L, Racoceanu D, Loménie N, Kulikova M, Irshad H, Klossa J, Capron F, Genestie C, Le Naour G, Gurcan MN. Mitosis detection in breast cancer histological images an ICPR 2012 contest. *J Pathol Inform*. 2013;4:8.
- Veta M, van Diest PJ, Willems SM, Wang H, Madabhushi A, Cruz-Roa A, et al. Assessment of algorithms for mitosis detection in breast cancer histopathology images. *Med Image Anal*. 2015;20(1):237–48.
- Racoceanu D, Capron F. Towards semantic-driven high-content image analysis: an operational instantiation for mitosis detection in digital histopathology. *Comput Med Imaging Graph*. 2015;42:2–15.
- Veta M, Heng YJ, Stathonikos N, Bejnordi BE, Beca F, Wollmann T, et al. Predicting breast tumor proliferation from whole-slide images: the TUPAC16 challenge. *Med Image Anal*. 2019;54:111–21.
- Balkenhol MCA, Bult P, Tellez D, Vreuls W, Claessen PC, Ciompi F, van der Laak JAWM. Deep learning and manual assessment show that the absolute mitotic count does not contain prognostic information in triple negative breast cancer. *Cell Oncol (Dordr)*. 2019;42(4):555–69.
- Tizooş HR, Pantanowitz L. Artificial intelligence and digital pathology: challenges and opportunities. *J Pathol Inform*. 2018;9:38.
- Metter DM, Colgan TJ, Leung ST, Timmons CF, Park JY. Trends in the US and Canadian pathologist workforces from 2007 to 2017. *JAMA Netw Open*. 2019;2(5):e194337.
- Maddox TM, Rumsfeld JS, Payne PRO. Questions for artificial intelligence in health care. *JAMA*. 2019;321(1):31–2.
- Stevanovic L, Choschick M, Moskovsky L, Varga Z. Variability of predictive markers (hormone receptors, Her2, Ki67) and intrinsic subtypes of breast cancer in four consecutive years 2015–2018. *J Cancer Res Clin Oncol*. 2019;145:2983–94.
- Paeng K, Hwang S, Park S, Kim M. A Unified Framework for Tumor Proliferation Score Prediction in Breast Histopathology. In: Cardoso M, et al., editors. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. DLMIA 2017, ML-CDS 2017. Lecture notes in computer science*, vol. 10553. Cham: Springer.
- Ren S, He K, Girshick R, Sun J. Faster r-cnn: Towards real-time object detection with region proposal networks. In: *Advances in neural information processing systems*; 2015. p. 91–9.
- He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*; 2016. p. 770–8.
- Tabata K, Uraoka N, Benhamida J, Hanna MG, Sirintrapan SJ, Gallas BD, et al. Validation of mitotic cell quantification via microscopy and multiple whole-slide scanners. *Diagn Pathol*. 2019;14(1):65.
- Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med*. 2019;17(1):195.
- Barinov L, Jairaj A, Becker M, Seymour S, Lee E, Schram A, Lane E, Goldszal A, Quigley D, Paster L. Impact of data presentation on physician performance utilizing artificial intelligence-based computer-aided diagnosis and decision support systems. *J Digit Imaging*. 2019;32(3):408–16.
- Chen J-M, Li Y, Xu J, Gong L, Wang L-W, Liu W-L, Liu J. Computer-aided prognosis on breast cancer with hematoxylin and eosin histopathology images: a review. *Tumor Biol*. 2017;39(3):1010428317694550.
- Kaman EJ, Smeulders AW, Verbeek PW, Young IT, Baak JP. Image processing for mitoses in sections of breast cancer: a feasibility study. *Cytometry*. 1984; 5(3):244–9.
- ten Kate TK, Beliën JA, Smeulders AW, Baak JP. Method for counting mitoses by image processing in Feulgen stained breast cancer sections. *Cytometry*. 1993;14(3):241–50.
- Wang H, Cruz-Roa A, Basavanahally A, Gilmore H, Shih N, Feldman M, Tomaszewski J, Gonzalez F, Madabhushi A. Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features. *J Med Imaging (Bellingham)*. 2014;1(3):034003.
- Veta M, van Diest PJ, Jiwa M, Al-Janabi S, Pluim JP. Mitosis counting in Breast Cancer: object-level Interobserver agreement and comparison to an automatic method. *PLoS One*. 2016;11(8):e0161286.

34. Beevi KS, Nair MS, Bindu GR. A multi-classifier system for automatic mitosis detection in Breast histopathology images using deep belief networks. *IEEE J Transl Eng Health Med*. 2017;5:4300211.
35. Nateghi R, Danyali H, Helfroush MS. Maximized inter-class weighted mean for fast and accurate mitosis cells detection in Breast Cancer histopathology images. *J Med Syst*. 2017;41(9):146.
36. Li C, Wang X, Liu W, Latecki LJ. DeepMitosis: mitosis detection via deep detection, verification and segmentation networks. *Med Image Anal*. 2018;45:121–33.
37. Puri M, Hoover SB, Hewitt SM, Wei BR, Adissu HA, Halsey CHC, et al. Automated computational detection, quantitation, and mapping of mitosis in whole-slide images for clinically actionable surgical pathology decision support. *J Pathol Inform*. 2019;10:4.
38. Hartman DJ, Van Der Laak JAWM, Gurcan MN, Pantanowitz L. Value of public challenges for the development of pathology deep learning algorithms. *J Pathol Inform*. 2020;11:7.
39. Wahab N, Khan A, Lee YS. Transfer learning based deep CNN for segmentation and detection of mitoses in breast cancer histopathological images. *Microscopy (Oxf)*. 2019;68(3):216–33.
40. Dessauvage BF, Thomas C, Robinson C, Frost FA, Harvey J, Sterrett GF. Validation of mitosis counting by automated phosphohistone H3 (PHH3) digital image analysis in a breast carcinoma tissue microarray. *Pathology*. 2015;47(4):329–34.
41. Focke CM, Finsterbusch K, Decker T, van Diest PJ. Performance of 4 Immunohistochemical Phosphohistone H3 antibodies for marking mitotic figures in Breast Cancer. *Appl Immunohistochem Mol Morphol*. 2018;26(1):20–6.
42. Davidson TM, Rendi MH, Frederick PD, Onega T, Allison KH, Mercan E, et al. Breast Cancer prognostic factors in the digital era: comparison of Nottingham grade using Whole slide images and glass slides. *J Pathol Inform*. 2019;10:11.
43. Norgan AP, Suman VJ, Brown CL, Flotte TJ, Mounajjed T. Comparison of a medical-grade monitor vs commercial off-the-shelf display for mitotic figure enumeration and small object (*helicobacter pylori*) detection. *Am J Clin Pathol*. 2018;149(2):181–5.
44. Bonert M, Tate AJ. Mitotic counts in breast cancer should be standardized with a uniform sample area. *Biomed Eng Online*. 2017;16:28.
45. Hanna M, Pantanowitz L. Redefining the high power field when counting mitoses using digital pathology. *Mod Pathol*. 2017;30(s2):396A.

Publisher's Note

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

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions



Pantanowitz, Liron

From: Parwani, Anil <Anil.Parwani@osumc.edu>
Sent: Sunday, August 30, 2020 9:57 PM
To: Pantanowitz, Liron
Subject: Re: Permission

Yes sure

Thanks

Anil Parwani

On Aug 30, 2020, at 12:56 PM, Pantanowitz, Liron <lironp@med.umich.edu> wrote:

CAUTION: External Email

Diagnostic Pathology
Editor-in-Chief: Anil V. Parwani

Dear Anil

RE: Permission to use Article for PhD Thesis

I recently published an article in *Diagnostic Pathology* 2020; 15:80 (attached as a PDF). I would like to use this article as part of my PhD thesis submitted at the University of Witwatersrand, Johannesburg, South Africa. Please can I request permission to do so. Thanks

Regards,

Liron Pantanowitz M.D.

Director, Division of Anatomic Pathology
 A. James French Professor of Pathology
 Department of Pathology & Clinical Labs
 NCRC Bldg 35, 2800 Plymouth Road
 University of Michigan, Ann Arbor, MI
 Tel 734-232-7318

lironp@med.umich.edu

<image001.png>

Electronic Mail is not secure, may not be read every day, and should not be used for urgent or sensitive issues

CHAPTER 4: PUBLISHED ARTICLE 3

Pantanowitz L, Quiroga-Garza GM, Bien L, Heled R, Laifenfeld D, Linhart C, Sandbank J, Albrecht Shach A, Shalev V, Vecsler M, Michelow P, Hazelhurst S, Dhir R. Clinical validation and deployment of an AI-based algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies. *Lancet Digital Health*, 2020; 2(8): e407-e416.

<https://pubmed.ncbi.nlm.nih.gov/33328045/>

Abstract

Background: There is high demand to develop computer-assisted diagnostic tools to evaluate prostate core needle biopsies (CNBs), but little clinical validation and a lack of clinical deployment of such tools. We report here on a blinded clinical validation study and deployment of an artificial intelligence (AI)-based algorithm in a pathology laboratory for routine clinical use to aid prostate diagnosis.

Methods: An AI-based algorithm was developed using haematoxylin and eosin (H&E)-stained slides of prostate CNBs digitised with a Philips scanner, which were divided into training (1 357 480 image patches from 549 H&E-stained slides) and internal test (2501 H&E-stained slides) datasets. The algorithm provided slide-level scores for probability of cancer, Gleason score 7-10 (vs Gleason score 6 or atypical small acinar proliferation [ASAP]), Gleason pattern 5, and perineural invasion and calculation of cancer percentage present in CNB material. The algorithm was subsequently validated on an external dataset of 100 consecutive cases (1627 H&E-stained slides) digitised on an Aperio AT2 scanner. In addition, the AI tool was implemented in a pathology laboratory within routine clinical workflow as a second read system to review all prostate CNBs. Algorithm performance was assessed with area under the receiver operating characteristic curve (AUC), specificity, and sensitivity, as well as Pearson's correlation coefficient (Pearson's r) for cancer percentage.

Findings: The algorithm achieved an AUC of 0·997 (95% CI 0·995 to 0·998) for cancer detection in the internal test set and 0·991 (0·979 to 1·00) in the external validation set. The AUC for distinguishing between a low-grade (Gleason score 6 or ASAP) and high-grade (Gleason score 7-10) cancer diagnosis was 0·941 (0·905 to 0·977) and the AUC for detecting Gleason pattern 5 was 0·971 (0·943 to 0·998) in the external validation set. Cancer percentage calculated by pathologists and the algorithm showed good agreement ($r=0·882$, 95% CI 0·834 to 0·915; $p<0·0001$) with a mean bias of -4·14% (-6·36 to -1·91). The algorithm achieved an AUC of 0·957 (0·930 to 0·985) for perineural invasion. In routine practice, the algorithm was used to assess 11 429 H&E-stained slides pertaining to 941 cases leading to 90 Gleason score 7-10 alerts and 560 cancer alerts. 51 (9%) cancer alerts led to additional cuts or stains being ordered, two (4%) of which led to a third opinion request. We report on the first case of missed cancer that was detected by the algorithm.

Interpretation: This study reports the successful development, external clinical validation, and deployment in clinical practice of an AI-based algorithm to accurately detect, grade, and evaluate clinically relevant findings in digitized slides of prostate CNBs.

An artificial intelligence algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies: a blinded clinical validation and deployment study

Liron Pantanowitz, Gabriela M Quiroga-Garza, Lilach Bien, Ronen Heled, Daphna Laifenfeld, Chaim Linhart, Judith Sandbank, Anat Albrecht Shach, Varda Shalev, Manuela Vecsler, Pamela Michelow, Scott Hazelhurst, Rajiv Dhir



Summary

Background There is high demand to develop computer-assisted diagnostic tools to evaluate prostate core needle biopsies (CNBs), but little clinical validation and a lack of clinical deployment of such tools. We report here on a blinded clinical validation study and deployment of an artificial intelligence (AI)-based algorithm in a pathology laboratory for routine clinical use to aid prostate diagnosis.

Methods An AI-based algorithm was developed using haematoxylin and eosin (H&E)-stained slides of prostate CNBs digitised with a Philips scanner, which were divided into training (1357480 image patches from 549 H&E-stained slides) and internal test (2501 H&E-stained slides) datasets. The algorithm provided slide-level scores for probability of cancer, Gleason score 7–10 (vs Gleason score 6 or atypical small acinar proliferation [ASAP]), Gleason pattern 5, and perineural invasion and calculation of cancer percentage present in CNB material. The algorithm was subsequently validated on an external dataset of 100 consecutive cases (1627 H&E-stained slides) digitised on an Aperio AT2 scanner. In addition, the AI tool was implemented in a pathology laboratory within routine clinical workflow as a second read system to review all prostate CNBs. Algorithm performance was assessed with area under the receiver operating characteristic curve (AUC), specificity, and sensitivity, as well as Pearson's correlation coefficient (Pearson's r) for cancer percentage.

Findings The algorithm achieved an AUC of 0.997 (95% CI 0.995 to 0.998) for cancer detection in the internal test set and 0.991 (0.979 to 1.00) in the external validation set. The AUC for distinguishing between a low-grade (Gleason score 6 or ASAP) and high-grade (Gleason score 7–10) cancer diagnosis was 0.941 (0.905 to 0.977) and the AUC for detecting Gleason pattern 5 was 0.971 (0.943 to 0.998) in the external validation set. Cancer percentage calculated by pathologists and the algorithm showed good agreement ($r=0.882$, 95% CI 0.834 to 0.915; $p<0.0001$) with a mean bias of -4.14% (-6.36 to -1.91). The algorithm achieved an AUC of 0.957 (0.930 to 0.985) for perineural invasion. In routine practice, the algorithm was used to assess 11429 H&E-stained slides pertaining to 941 cases leading to 90 Gleason score 7–10 alerts and 560 cancer alerts. 51 (9%) cancer alerts led to additional cuts or stains being ordered, two (4%) of which led to a third opinion request. We report on the first case of missed cancer that was detected by the algorithm.

Interpretation This study reports the successful development, external clinical validation, and deployment in clinical practice of an AI-based algorithm to accurately detect, grade, and evaluate clinically relevant findings in digitised slides of prostate CNBs.

Funding Ibex Medical Analytics.

Copyright © 2020 The Author(s). Published by Elsevier Ltd. This is an Open Access article under the CC BY-NC-ND 4.0 license.

Introduction

Adenocarcinoma of the prostate is the second most common cancer diagnosed in men,¹ with more than 1 million newly diagnosed cases of prostate cancer annually. Hence, prostate specimens are frequently encountered in surgical pathology practice. Today, the histopathological assessment of biopsy tissue is the mainstay of diagnosing prostate cancer, which includes core needle biopsy (CNB) and, if warranted, surgical resection. For most pathology laboratories, the current method of rendering a tissue diagnosis involves light microscopic

examination of haematoxylin and eosin (H&E)-stained tissue sections.² Due to changing guidelines, there has been a dramatic increase in the number of CNBs reviewed per case over the past decade. Coupled with an increase in overall cancer incidence and growing shortage of pathologists worldwide,³ there is an emerging need to develop automated, artificial intelligence (AI)-based tools to support pathologists. Management of afflicted men hinges on diagnosis and on the Gleason grade of their prostate adenocarcinoma for disease prognosis. Reliably diagnosing and correctly grading prostate adenocarcinoma in CNBs is

Lancet Digital Health 2020; 2: e407–16

See [Comment](#) page e383

Department of Pathology, University of Pittsburgh Medical Center, Pittsburgh, PA, USA (L Pantanowitz MD, G M Quiroga-Garza MD, R Dhir MD); Department of Anatomical Pathology, University of the Witwatersrand and National Health Laboratory Services, Johannesburg, South Africa (L Pantanowitz, P Michelow MBBCh); Ibex Medical Analytics, Tel Aviv, Israel (L Bien, R Heled, D Laifenfeld PhD, C Linhart PhD, J Sandbank MD, M Vecsler PhD); Institute of Pathology, Maccabi Healthcare Services, Rehovot, Israel (J Sandbank); Shamir Medical Center, Beer Yaakov, Israel (A Albrecht Shach MD); KSM Research and Innovation Institute, Maccabi Healthcare Services, Tel Aviv, Israel (V Shalev MD); and School of Electrical & Information Engineering and Sydney Brenner Institute for Molecular Bioscience, University of the Witwatersrand, Johannesburg, South Africa (S Hazelhurst PhD)

Correspondence to: Dr Liron Pantanowitz, Department of Pathology, UPMC Cancer Pavilion, Pittsburgh, PA 15232, USA

lironp@med.umich.edu

Research in context

Evidence before this study

We searched PubMed and Google Scholar on Feb 20, 2020, with the keywords “artificial intelligence” OR “deep learning” OR “machine learning” AND “pathology” AND “prostate cancer”. The advanced search was limited to the English language. This search rendered 185 results, but only a few relevant studies retrieved employed image analysis for cancer detection or artificial intelligence (AI)-based Gleason grading performed on whole slide images. Additional relevant articles were reviewed from the publications’ references. Despite much discussion in the field, there were few studies actually validating high performance of these algorithms, especially in large independent blinded studies. Most studies report performance on their internal test sets and only on few external validation sets. In the context of supporting clinical decisions, the need for an AI-based tool that combines high accuracy levels validated on large independent cohorts that incorporates clinically meaningful features is still unmet. Furthermore, the technical feasibility of deploying such a system in routine clinical workflow has not been addressed before. To the best of our knowledge, no report regarding clinical deployment of an AI system embedded into routine pathology practice has yet been published.

Added value of this study

We report the validation and performance of an AI-based prostate histopathology algorithm that extends beyond just cancer detection and grading, but also measures cancer proportion and the detection of perineural invasion to meet clinical reporting needs. Additionally, we describe the clinical deployment of such an algorithm in routine clinical practice, where implementation of a second read system showed early utility in preventing a missed prostate cancer diagnosis.

Implications of all the available evidence

An accurate, robust, and validated AI-based algorithm to detect, grade, and automatically impart clinically relevant diagnostic parameters regarding prostate adenocarcinoma offers an important tool for computer-assisted diagnosis in routine pathology practice. Demonstration of the technical feasibility for deploying such an AI-based system in the routine clinical workflow of a pathology laboratory elevates existing discussion of AI-based tools in pathology to a practical level, revealing how computational pathology can lead to improved efficiency, accuracy, consistent diagnoses, and better patient management.

challenging because of potential cancer mimics and the presence of only small foci of well differentiated adenocarcinoma. Hence, for an automated tool to be adopted into practice, it will need to have clinical grade accuracy.

Recent technological and regulatory advances in whole slide imaging have accelerated the adoption of digital imaging in pathology.⁴ Now that pathology departments have several commercial digital pathology platforms available for diagnostic work, there is interest in leveraging AI tools. Several publications have shown the feasibility of developing AI-based algorithms to analyse histopathology images, specifically for prostate cancer.^{4–8} However, there is minimal literature clinically validating such algorithms using large, independent blinded studies. Furthermore, routine clinical deployment of such deep learning systems in pathology laboratories has not been reported.

The aim of this study was twofold. The first aim was to do a blinded clinical validation of an AI-based algorithm that extended beyond just prostate adenocarcinoma detection and Gleason grading, by also detecting clinically meaningful features such as tumour extent and perineural invasion. The second aim was to deploy this system in a pathology laboratory for routine clinical use.

Methods

Study design

The study encompasses three key steps: (1) development and testing of an AI-based algorithm for prostate CNBs; (2) blinded algorithm validation in an external, independent

dataset; and (3) algorithm deployment in routine clinical use. Each of these steps is described below.

Institutional review board approval was obtained for this study (University of Pittsburgh Institutional Review Board PRO18030439; Maccabi Ethics Helsinki Committee 0153-16-ASMC and 0081-18-BBL; and University of the Witwatersrand, Johannesburg Human Research Ethics Committee [Medical] M191003).

Algorithm development

Training and internal test datasets came from prostate CNBs retrieved from the archive of the Pathology Institute at Maccabi Healthcare Services’ centralised laboratory (MegaLab) in Israel. H&E-stained slides were scanned using a Philips IntelliSite Scanner (Philips Digital Pathology Solutions; Best, Netherlands) at 40× magnification (resolution of 0·25 µm/pixel).

The algorithm that we developed, whose core technology is based on multilayered convolutional neural networks (CNNs) that were specifically designed for image classification tasks, analyses a whole slide image in three consecutive steps: tissue detection, classification, and slide-level analysis. Briefly, the first step uses a Gradient Boosting classifier, trained on thousands of image patches, to distinguish between tissue and background areas within the slide. After this, an ensemble of three CNN-based models is run on all tissue areas.

The models were trained on 1357480 labelled image patches that were extracted from manual annotations on 549 slides, selected from more than 65 000 slides in the

archive on the basis of various criteria, such as reported Gleason grade and rare findings. Annotations were done by three senior pathologists, each with 20–40 years of experience. The classification step yields predictions (probabilities) for every area in the slide belonging to each of 18 predefined classes, such as Gleason pattern 3 cancer, normal gland, and chronic inflammation (appendix p 6). Finally, the third step combines the 18 probability heatmaps (appendix p 6) and calculates slide-level scores for chosen endpoints.

Low concordance in Gleason scoring between pathologists' confounds the ability to obtain robust ground truth. Therefore, we focused on groupings of clinical significance rather than the full range of individual Gleason scores, choosing the following five endpoints: presence of cancer (primary endpoint), Gleason score 7–10 (including scores 3+4, 4+3, 4+4, 3+5, 5+3, 4+5, 5+4, and 5+5) versus Gleason score 6 (3+3) or atypical small acinar proliferation (ASAP), Gleason pattern 5 (including scores 3+5, 5+3, 4+5, 5+4, and 5+5), perineural invasion, and calculation of the percentage of cancer present in CNB material. The Gleason endpoints represent clinically relevant endpoints for disease management (Gleason score 7–10) and aggressive cancers (Gleason pattern 5).⁹ More details on algorithm development are provided in the appendix (pp 2–3).

Algorithm testing

To evaluate the accuracy of the prostate algorithm, its output was compared with ground truth on a large collection of whole slide images independent from the training set. The internal test dataset included all 213 consecutive prostate CNBs (2576 H&E-stained slides) received at MegaLab from March 1 to June 30, 2016, which were also digitised with a Philips scanner. Associated immunohistochemistry slides from these cases were used to review cases and establish ground truth. 51 (2.0%) H&E-stained slides were not scanned due to limiting slide physical conditions (eg, broken glass or missing slides), and 24 (0.9%) H&E-stained slides that had been chosen previously for algorithm training as part of the randomly selected CNBs were filtered out. Thus, the internal test was run on 2501 H&E-stained slides from 210 cases. Ground truth was established at the slide level on the basis of the pathologists' diagnosis, either from the original pathology report (for benign cases) or after review by two senior pathologists with 40 years (JS) and 20 years (AAS) of experience for cases where there was a diagnosis of cancer in the report.

External validation

External validation of the algorithm was done at the University of Pittsburgh Medical Center (UPMC) using a validation dataset with different pre-imaging and scanning parameters. Glass slides at UPMC were scanned at 40x magnification (0.25 µm/pixel resolution) using an Aperio AT2 scanner (Leica Biosystems; Buffalo Grove, IL,

USA). UPMC diagnoses are reported per part, where a part is one of the three biopsy regions (upper, mid, or base) in one of the bilateral prostate lobes. One part is often represented in multiple whole slide images.

To prepare the algorithm for external validation, we used a set of 32 prostate CNB cases (selected from cases occurring between August, 2014, and January, 2018), comprising 159 parts, to calibrate the algorithm for UPMC-specific whole slide image attributes (eg, scanner and staining) and to verify the technical validity of the whole slide images (eg, file format and resolution). This set included diagnoses with cancers of various Gleason scores (appendix p 5), high-grade prostatic intraepithelial neoplasia, inflammation, and atrophy. We divided the set into a calibration set (also known as a tuning set) and an internal test set. The calibration set comprised 44 parts, selected according to criteria used in previous training slides (appendix pp 2–3), that were manually annotated by senior pathologists (JS, AAS). The remaining 115 parts that were not annotated were used for internal validation.

A separate archival dataset of whole slide images that included 100 consecutive cases of prostate CNB cases previously received and signed out (ie, formally diagnosed and reported) by the Genitourinary Center of Excellence at UPMC served as an independent, external, blinded validation set. Each case included all associated H&E-stained slides, giving a total of 1627 H&E-stained whole slide images. The 100 cases were organised into 379 parts with an average of four slides per part (typically deeper H&E-stained section levels of the same cores). All study data were anonymised and metadata for each case (eg, patient demographics and immunohistochemistry results) were recorded. Patients' age distribution, percentage of cancer cases, and Gleason score distribution are summarised in the appendix (p 7).

The algorithm was applied to the external, blinded validation set. Data were subsequently unblinded and the results of the algorithm compared with the ground truth diagnoses. Assessment of performance was done on a per-part level. The performance on detection of the five endpoints was assessed.

Ground truth

Ground truth was established on the basis of the original UPMC pathology report using the diagnosis rendered by one of three genitourinary subspecialty pathologists before this study. To address the potential for discordance among pathologists, and to ensure accurate assessment of algorithmic results, a subset of the parts in the external validation set were chosen for ground truth ascertainment. Ground truth ascertainment was done on four of the features assessed in the external validation set (ie, cancer detection, Gleason score 7–10, Gleason pattern 5, and perineural invasion); cancer percentage was not subject to discrepancy analysis owing to the protocol-specific nature of discrepancies. The subset of parts selected for ground truth ascertainment was based

See Online for appendix

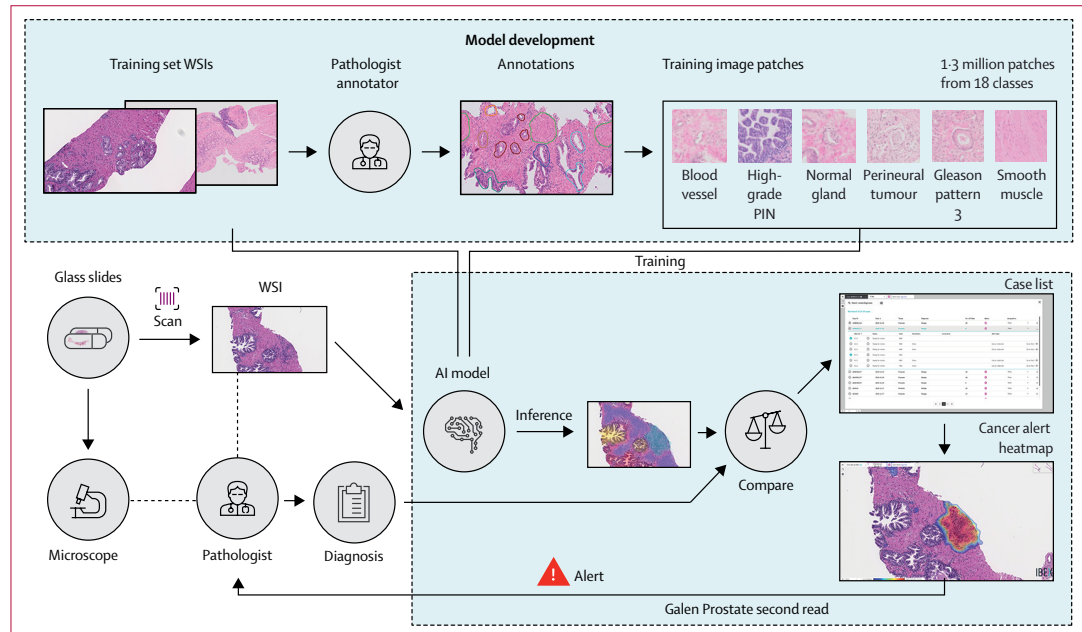


Figure 1: Overview of the algorithm and clinical deployment of the Galen Prostate second read system
AI=artificial intelligence. WSI=whole image slide. PIN=prostatic intraepithelial neoplasia.

on likelihood of affecting the performance metrics of the algorithm. Thus, 10% of the highest-scoring negative parts (eg, parts reported as benign in the pathology report) and 5% of the lowest-scoring positive parts (eg, parts reported as positive for cancer in the pathology report) in the external validation set were sent for review by two independent pathologists with genitourinary pathology expertise who did not previously diagnose any of the cases, one from UPMC (GMQ-G), and one from Maccabi (JS). Their review was done in a blinded manner using the same digital slides and image viewer to ensure uniformity between reviewers. Ground truth was based on consensus between the two pathologists. For parts where there was an initial discrepancy between the two pathologists, a third pathologist (RD) with genitourinary pathology expertise blindly reviewed the slides and the majority vote was used as the final consensus diagnosis. 65 parts were reviewed (appendix p 7).

Deployment for routine clinical use: second read application

Maccabi Healthcare Services is a large health-care provider in Israel with a centralised pathology institute with approximately 120 000 surgical pathology cases annually, including around 700 prostate CNB cases reviewed by three genitourinary subspecialty pathologists out of a team of 12 senior pathologists. Around 40% of these CNBs are diagnosed with cancer. Galen Prostate (Ibex Medical Analytics), the product based on the prostate algorithm, has been implemented in Maccabi Pathology

Institute since March, 2018, as a second read system—namely, a quality control application that reviews whole slide images of all prostate CNBs (figure 1). Slides were scanned using a Philips IntelliSite Scanner and processed via a server with four graphics processing units. With this setup, the algorithm was run in parallel to the pathologists' routine workflow. Triggers were used to alert pathologists in the event there was a case with a discrepancy between their diagnosis and the output of the algorithm. The system generated alerts for slides with a high algorithmic cancer score that were diagnosed as benign by the pathologist, and for slides with a high Gleason score (ie, 7–10) that were diagnosed as Gleason score 6 by the pathologist. The alert threshold was set to correspond to a specificity of 90% (see appendix pp 3–4 for threshold rationale). Pathologists were able to view the case list with alerts and do a second review, specifically focused on the region that triggered the alert outlined by heatmaps in the second read slide viewer. The system took 1 month to set up, including automatic export of slides from the Philips Image Management System to Galen Prostate. A training session was held for each user and for the technician scanning the slides.

Statistical analysis

Sample size for 80% power with two-sided 5% level of significance was calculated on the basis of test performance (appendix p 4). Area under the receiver operating characteristic curve (AUC) was calculated with 95% Wald CIs using the continuous score. Specificity

	Number of slides (internal test) and parts (external validation)	AUC (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	PPV	NPV
Internal test (Maccabi Healthcare Services)						
Benign vs cancer	2501: 1957 benign and 490 cancer*	0.997 (0.995–0.998)	90.14% (87.76–92.09)†	99.59% (98.39–99.90)	71.7%	99.9%
External validation (UPMC)						
Benign vs cancer	355: 225 benign and 130 cancer‡	0.991 (0.979–1.00)	97.33% (94.43–98.74)	98.46% (94.06–99.61)	95.5%	99.1%
Gleason score 6 or ASAP vs Gleason score 7–10§	151: 73 Gleason score 6 or ASAP and 78 Gleason score 7–10	0.941 (0.905–0.977)	90.41% (78.92–95.96)	85.9% (75.72–92.25)	90.5%	85.7%
ASAP or Gleason pattern 3 or 4 vs Gleason pattern 5¶	151: 131 ASAP or Gleason pattern 3 or 4 and 20 Gleason pattern 5	0.971 (0.943–0.998)	90.84% (84.18–94.87)	85% (51.24–96.83)	58.6%	97.5%
Cancer without vs with perineural invasion	154: 108 without perineural invasion and 46 with perineural invasion	0.957 (0.930–0.985)	90.74% (83.10–95.13)	86.96% (74.47–93.84%)	80%	94.2%

AUC=area under the receiver operating characteristic curve. PPV=positive predictive value. NPV=negative predictive value. UPMC=University of Pittsburgh Medical Center. ASAP=atypical small acinar proliferation. *34 slides with ASAP and 20 slides for which pathologists did not reach a conclusion were excluded. †Selected to reflect pathologists' review using 10% of slides for quality control. ‡Parts that were diagnosed as ASAP by the pathologists were excluded. §Gleason score 7–10 includes Gleason scores 3+4, 4+3, 4+4, 3+5, 5+3, 4+5, 5+4, and 5+5. ¶ASAP or Gleason pattern 3–4 refers to parts not having Gleason pattern 5—ie, diagnosed as ASAP or Gleason score 3+3, 3+4, 4+3, or 4+4. Gleason pattern 5 includes Gleason scores 3+5, 5+3, 4+5, 5+4, and 5+5. ||154 parts include: 151 parts with adenocarcinoma and Gleason score, two parts with other cancer, and one part where a consensus was not reached about the Gleason score during ground truth ascertainment.

Table 1: Algorithm performance

and sensitivity at multiple cutoff points were calculated for each feature and are presented with two-sided 95% CIs. The CIs were calculated from a generalised estimating equation model using the GENMOD procedure in SAS (dist=bin) using the ilink option in the Lsmmeans statement. This was done to accommodate the within-subject correlation due to repeated measurements. Pearson's correlation coefficient is presented for cancer percentage concordance between the algorithm and pathology report. The mean bias (difference between cancer percentage) and its SD are presented as further measures of concordance, as well as the Bland-Altman 95% limits of agreement (presented with 95% CIs). We used SAS version 9.4 for all analyses.

Role of the funding source

The funder of the study supported study design and writing of the report. LP had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Results

Patient characteristics in the study cohorts are shown in the appendix (pp 7, 11) and reflect the representative population undergoing prostate CNBs in Israel and the USA (appendix p 3).

The AUC for cancer detection was 0.997 (95% CI 0.995–0.998) on the internal test set (appendix p 7) and 0.991 (0.979–1.00) on the UPMC external validation set (appendix pp 8, 11; table 1). Additional performance metrics in internal test and external validation sets are shown in table 1.

Ground truth ascertainment for cancer versus benign diagnosis resulted in algorithm-driven diagnostic corrections for seven parts in the external validation dataset:

four parts were corrected from benign to cancer (two parts) or ASAP (two parts; appendix p 13) and three parts were corrected from cancer to benign (one part) or ASAP (two parts). At the case level, one case that was originally diagnosed as ASAP was changed to cancer (figure 2B).

Cancer percentage correlation was computed on 126 parts diagnosed as cancer in UPMC pathology reports. Figure 2A shows an example of the algorithm results. There was high correlation between the cancer percentage reported by pathologists with the percentage computed by the algorithm (Pearson's $r=0.882$, 95% CI 0.834 to 0.915; $p<0.0001$; appendix p 12). The mean bias was -4.14% (95% CI -6.36 to -1.91), showing that, on average, the algorithm underestimates the cancer percentage by 1.9–6.3% (appendix p 8). The Bland-Altman 95% limits of agreement, which show the difference between reported and calculated percentage, are -28.6% (95% CI -33.9 to -23.3 ; algorithm lower) to 20.3% (15.0 to 25.7; algorithm higher). Examining discrepancies between the algorithm and the pathologist in cancer percentage reveals that these variances stem largely from specific calculation protocols, rather than algorithmic inaccuracies—eg, the decision to include small benign areas between cancer foci in the calculation (appendix p 14).

When assessing the algorithm's performance on Gleason grading, the AUC for distinguishing between ASAP or Gleason score 6 versus higher Gleason scores was 0.941 (95% CI 0.905–0.977) and 0.971 (0.943–0.998) for detecting any Gleason pattern 5 in a CNB (table 1). Additional performance metrics on grading performance are summarised in table 1 and in the appendix (p 9).

Ground truth ascertainment resulted in algorithm-driven corrections of four parts from Gleason score 6 to Gleason score 7–10 (figure 2C), and two parts from Gleason

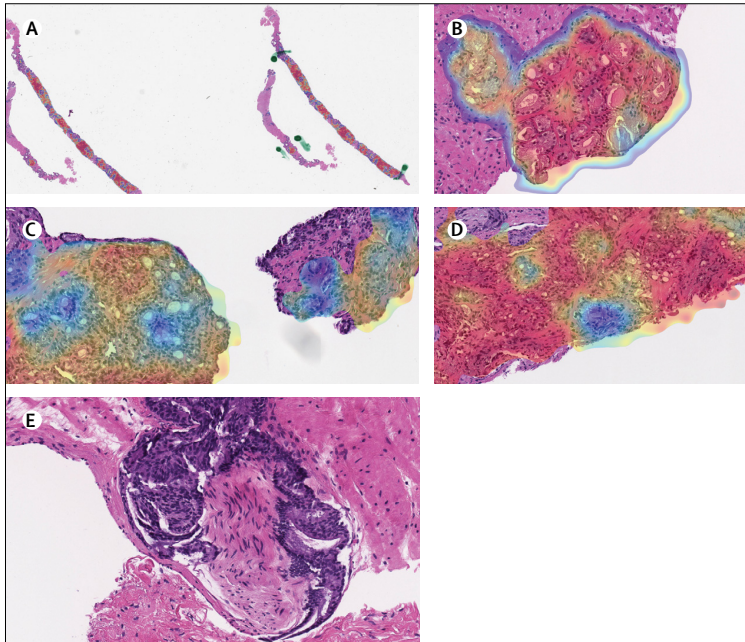


Figure 2: Examples of diagnoses after review

All images are stained with haematoxylin and eosin and displayed at $0.6\times$ ($16.67\ \mu\text{m}/\text{pixel}$; panel A) and $20\times$ ($0.50\ \mu\text{m}/\text{pixel}$; panels B–E) magnification. (A) Example of cancer proportion reported by pathologists versus calculations performed by the algorithm in the external validation dataset. Prostate CNBs (part 662_6) with the cancer heatmap (where blue shows low probability and red shows high probability) displayed. Tumour proportion calculated by the algorithm was 33% compared with 40% reported by a pathologist. Panels B–E provide examples of revised diagnoses. (B) Prostate CNB (part 665_1) with cancer heatmap (where blue shows low probability and red shows high probability) originally diagnosed as benign that was subsequently changed to cancer with Gleason score $6+3$ after review. The case diagnosis was updated from ASAP to cancer. (C) Prostate biopsy (part 598_5) with a Gleason pattern heatmap (where blue shows Gleason pattern 3, yellow shows Gleason pattern 4, and red shows Gleason pattern 5) that was updated from Gleason score $6+3$ to Gleason score 7 (there was no concordance between reviewers as to $3+4$ vs $4+3$). (D) Prostate biopsy (part 606_1) with a Gleason pattern heatmap (where blue shows Gleason pattern 3, yellow shows Gleason pattern 4, and red shows Gleason pattern 5). This biopsy part and case were updated from Gleason score $7+3$ to Gleason score $8+3$. (E) Example of a prostate CNB (part 686_4) that was updated to positive for perineural invasion after review. The colours represent the original staining. CNB=core needle biopsy. ASAP=atypical small acinar proliferation.

	False positive	False negative
Adenocarcinoma	1 benign, 2 ASAP	2 cancer, 2 ASAP
Gleason score 7–10	2	4
Gleason pattern 5	0	1
Perineural invasion	1	2

Data are number of parts. ASAP=atypical small acinar proliferation.

Table 2: Pathologists' misdiagnoses identified by the algorithm

score 7–10 to Gleason score 6. One part was corrected from Gleason score 7 to Gleason grading with pattern 5 present (figure 2D).

The algorithm detected perineural invasion with an AUC of 0.957 (95% CI 0.930–0.985), further summarised in table 1 and in the appendix (p 9).

Ground truth ascertainment resulted in algorithm-driven corrections for two parts from negative to positive for perineural invasion (figure 2E). For one part that was

positive for perineural invasion in the report, it was determined that perineural invasion was absent.

Discrepancies between the two reviewing pathologists for 65 parts that were sent for ground truth ascertainment were observed for all endpoints analysed, with major discrepancies in Gleason grading (nine [14%] parts for Gleason grade group difference >1 and a total of 24 [37%] parts with differences in Gleason grading) and perineural invasion diagnosis (nine [14%] parts), followed by distinguishing between benign tissue and cancer (six [9%] parts) and between ASAP and cancer (two [3%] parts). Misdiagnoses by pathologists occurred for 17 parts and six cases, including one case in which cancer was misdiagnosed as ASAP and five cases where the revised diagnosis changed the Gleason score, with potential clinical significance (table 2).

Between March, 2018, and November, 2019, 941 cases (11429 H&E-stained slides) were processed by the second read system at the Maccabi Pathology Institute. Cancer alerts were raised for an average of 10.9% of the slides belonging to cases diagnosed by pathologists as benign, with 1.35 cancer alerts (slides) per such case. Of 1451 slides from 115 cases diagnosed as Gleason score $6+3$, the system issued 90 (6.2%) Gleason score 7–10 alerts and 560 (38.6%) cancer alerts. Upon pathologist review of the cancer alerts, 509 (90.9%) alerts required no substantial effort: most areas that triggered the alerts were identified as either atrophic glands or crushed glands, consistent with an irregular feature that mimics malignant glands. 51 (9.1%) alerts led to additional cuts or stains being ordered, two (4%) of which led to a third opinion request. Alerts were focused on specific areas, meaning that review time was minimal, resulting overall in approximately 1% of the pathologist's time.

We report on the first case with missed cancer in a CNB detected by the system, which occurred immediately following deployment. The biopsy from a 55-year-old patient was diagnosed by the pathologist as benign with foci of acute and chronic inflammation and atrophic changes in both prostate gland lobes. The second read system raised alerts for the presence of cancer in three slides within the case with a high probability score for cancer (>0.99 ; figure 3). Subsequent re-examination by the pathologist, additional recut sections, and an immunohistochemistry stain for CK903 were done, and the diagnosis was revised to minute foci of adenocarcinoma of acinar type with Gleason score 6 in the right lobe core biopsy. Three cores were reported as involved, and the estimated percentage of tumour in prostatic tissue was less than 5%. The left core biopsy diagnosis remained unchanged (ie, negative). As a result, this patient was included in an active surveillance protocol and his prostate-specific antigen level taken 3 months after this biopsy was above normal at 5.56 ng/mL and continued to rise during the surveillance to 7.93 ng/mL. Additional missed cancers were identified during the clinical deployment, details of which were not provided by the laboratory.

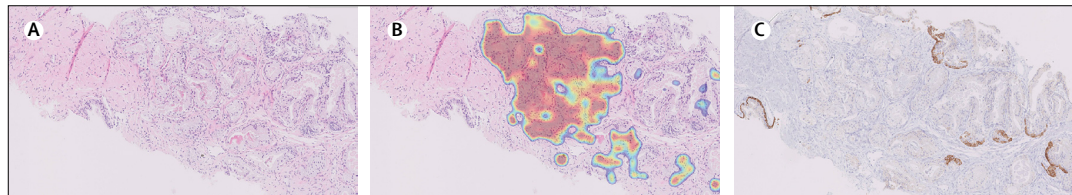


Figure 3: Missed cancer case originally diagnosed as benign

Images in panels A and B are stained with haematoxylin and eosin. All images are displayed at 10× (1 µm/pixel) magnification. (A) Prostate biopsy showing abnormal glandular focus. (B) Algorithm heatmap detecting cancer with high probability (red areas) in the same small focus. (C) Corresponding area on the immunohistochemistry slide (CK903) with absent basal cells confirming the diagnosis of adenocarcinoma.

Discussion

We report the development of a medical-grade AI-based algorithm for evaluating digitised prostate CNB slides, and successful deployment of this AI tool in routine clinical practice. We show high accuracy of the algorithm, based on a large, blinded, external validation dataset, to identify and quantify prostate cancer, differentiate between low-grade and high-grade tumours, and detect perineural invasion in CNBs. To the best of our knowledge, this is the first report of an AI-based algorithm that extends beyond cancer detection and grading of prostatic cancer in histopathological images, and one of the first instances of clinical use of an AI-based algorithm in routine pathology practice. We are aware of one other study where implementation of an AI-based algorithm for the automatic detection of prostate cancer in a pathology laboratory was reported,¹⁰ but no outcomes were published.

AI algorithms and models are typically developed from data that are assumed to be representative. Overfitting is a common problem, where despite good performance on training data or on test data similar to the training data, performance deteriorates on novel data. Multiple strategies exist to avoid overfitting (eg, cross-validation or bootstrapping), but true performance can only be determined through blinded studies with external datasets. The demonstration herein of high performance of the algorithm using an external, geographically distinct, blinded dataset is crucial for true evaluation of the performance and utility of the algorithm. To our knowledge, very few studies have attempted such validation, with narrower applications and significantly lower performance.

Deployment of such an AI tool in clinical practice is timely, not only because prostate adenocarcinoma is one of the most common cancers seen in men, but also due to the substantial increase in pathologists' workload as cancer cases rise, combined with increased complexity of histopathological assessment with changes in guideline recommendations. Unfortunately, there is a concomitant decline in the pathology workforce. When adjusted by new cancer cases per year, the workload per US pathologist has risen by 41.73%¹ and this gap is growing, potentially resulting in delayed cancer diagnoses and diagnostic errors. Reports of missed prostate cancer and

a lack of concordance with Gleason grading have been well documented,¹¹ especially when these diagnoses are rendered by general pathologists instead of subspecialist urological pathologists.^{12–15} Raciti and colleagues¹⁶ showed that in a non-clinical setting, AI can increase the detection of prostate cancer in whole slide images of CNBs. In our study, the two pathologists who participated disagreed on cancer detection in 9% of biopsy parts and on Gleason grade grouping in 37% of biopsy parts (14% major disagreement). Overall, 17 misdiagnosed parts and six misdiagnosed cases in the UPMC dataset were identified in this study, including misdiagnoses in cancer detection and grading, as well as detection of perineural invasion, some of which might have affected treatment. In a pathology laboratory in France, the same second read system identified 12 misdiagnosed cases, including missed high-grade cancers.¹⁷ Several pathology practices have established a second review process for quality control purposes on a portion (eg, 10%) of cases to thwart misdiagnoses.¹⁸ Although useful, this additional quality control step, when done manually, further adds to a pathologist's workload, and therefore is typically practised rarely. AI tools, as shown in this study, can be leveraged to help to automate this safety net task.^{19,20} Indeed, the deployment described here of Galen Prostate as a second read application is the first instance of 100% quality control of prostate CNBs in a laboratory, drastically diminishing the chances for misdiagnoses with negligible impact on a pathologist's workload.

Previous publications of prostate cancer algorithms devoted to analysing histopathology images have reported lower or similar performance characteristics for cancer detection to those obtained herein (table 3).^{6,8,21–25} However, many previous publications on this topic report on algorithms limited to just one task (ie, narrow AI) and provide performance metrics verified mostly on internal test sets. For example, several researchers have explored the use of machine and deep learning techniques to provide just Gleason grading.^{23,24,26,27} By contrast, our study reports high performance characteristics of a multifunction algorithm for prostate CNB interpretation (ie, cancer detection, grading, evaluation of tumour extent, and perineural invasion), assessed on an independent external dataset. The algorithm in our study was able to simultaneously evaluate CNBs for

	Dataset dimension	Cancer vs benign accuracy	Gleason grading accuracy
Performance assessed on external validation datasets			
Campanella et al (2019) ⁶	12 727 slides	AUC 0.932	Not available
Bulten et al (2020) ⁸	245 tissue microarray cores*	AUC 0.98	AUC 0.87 for benign or Gleason grade group 1 vs Gleason grade group 2–5 AUC 0.86 for benign or Gleason grade group 1–2 vs Gleason grade group 3–5
Nir et al (2019) ⁷	230 slides from 56 prostatectomy cases	AUC 0.75	Not available
Ström et al (2020) ²¹	330 cores from 73 cases	AUC 0.986	Cohen's κ 0.7
Current study	1627 slides from 100 cases	AUC 0.991 overall AUC 0.957 for perineural invasion†	AUC 0.941 for ASAP or Gleason score 6 vs Gleason score 7–10 AUC 0.971 for ASAP or Gleason pattern 3–4 vs Gleason pattern 5
Performance assessed only on internal test sets			
Arvaniti et al (2018) ³²	245 tissue microarray cores*	Not available	Cohen's κ 0.55
Nagpal et al (2019) ³³	1226 slides	Not available	AUC 0.7
Nir et al (2018) ²⁴	86 tissue microarray cores from 60 cases	AUC 0.85; sensitivity 91.3%; specificity 84%	Accuracy 79%; sensitivity 75.9%; specificity 77.9%
Lucas et al (2019) ²⁵	96 sections from 38 cases	Accuracy 92%; sensitivity 90%; specificity 93%	Accuracy 90%; sensitivity 77%; specificity 94%
Details on Gleason patterns, scores, and grading groups can be found in Rice-Stitt et al. ⁹ AUC=area under the receiver operating characteristic curve. ASAP=atypical small acinar proliferation. *The same tissue microarray was used in these studies. †The current study is the only study to assess perineural invasion.			
Table 3: Performance of algorithms in detection and grading of prostate cancer			

perineural invasion, a feature not reported by other researchers but one that is well known to have clinical and prognostic significance.²⁸ Notably, perineural invasion does not lend itself to some of the published AI-based algorithms, as it is typically very small (unlike adenocarcinoma, which sometimes covers a large fraction of the tissue core) and relatively uncommon. Furthermore, cancer percentage in CNBs calculated by our algorithm highly correlated with the proportion that was estimated by pathologists and can be further tailored to mimic the protocol for calculating cancer percentage for each laboratory. Ström and colleagues²¹ reported slightly lower correlation between their algorithm and pathologists when calculating tumour length in prostate biopsies compared with our algorithm for cancer percentage. This is not surprising, as the quantitative determination of prostate cancer (eg, tumour volume) in CNBs is controversial and fraught with subjectivity between pathologists.²⁹ Algorithms that can take these measurements offer laboratories an opportunity to harmonise clinical practice among pathologists.

Our study has several limitations. Gleason grading for prostate cancer is a well established grading system globally adopted for the treatment and prognosis of prostate cancer, and is considered the gold standard in the field. Still, discordance between pathologists in Gleason grading is widely acknowledged,¹¹ and thus a true and agreed-upon ground truth is sometimes difficult to reach. The AI-based algorithm employed in this study was trained using annotations from multiple independent pathologists and thus the computed

Gleason score of the deep-learning algorithm is expected to be generalisable, consistent with the high performance shown here. However, true and objective performance metrics can ultimately only be determined using a combination of a committee of expert pathologists for grading and long-term outcome data, which would also allow for individual grade assessment rather than the groupings used here. Ongoing studies are assessing these questions and are also focused on assessing performance characteristics of additional features computed by the algorithm, including areas corresponding to certain Gleason scores or percentage of certain Gleason scores within cancerous tissue, detection of high-grade prostatic intraepithelial neoplasia, and inflammation. These, as well as the performance of the algorithm in detection of rare variants of adenocarcinoma, are important features of the algorithm that were beyond the scope of the current study. Still, the utility of the study for wider deployment is evident, as we report here on six patient-level misdiagnoses at UPMC, including clinically significant misdiagnoses such as involvement of higher Gleason patterns.

Implementation of AI-based tools for routine clinical work carries practical considerations of generalisability, infrastructure requirements, and throughput. Although algorithms had been shown to have low performance for whole slide images of different file formats acquired by different scanners,³⁰ we were able to show equally high performance when slides were scanned using a Philips (clinical deployment dataset) and Leica (external validation dataset) scanner, although more work, such as

establishing a standardised file format—eg, Digital Imaging and Communications in Medicine—is warranted. We also show here that the algorithm calibration to a new laboratory requires a set of only around 30 cases. In fact, future versions of the algorithm incorporating additional training sets are underway, including a total of four different scanners and ten different laboratories, ultimately eliminating the need for calibration altogether. Finally, our approach to annotations enabled us to reach high accuracy with only three CNNs, as opposed to other published studies using a large number of CNNs, rendering their deployment impractical in terms of hardware and run-time requirements. From a throughput perspective, the system deployed in Maccabi analyses many slides per hour, keeping up with the rate at which slides are scanned.

In summary, we report the development, external clinical validation, and deployment in routine practice of an AI-based algorithm to detect, grade, and evaluate additional clinically relevant tumour features in digitised slides of prostate CNBs. These data suggest that this AI-based algorithm could be used as a tool to automate screening of prostate CNBs for primary diagnosis, assess signed-out cases for quality control purposes, and standardise reporting to improve patient management. Studies reporting on deployment in additional laboratories and associated clinical utility are underway.

Contributors

LP conceived the study. DL, CL, PM, and SH contributed to study design. LP, DL, CL, and RD contributed to the methodology. LP contributed to the project administration. LP and RD contributed to study coordination. LP provided informatics support. LP, LB, DL, and MV contributed to the literature search. LP, LB, RH, CL, JS, VS, and RD contributed to data collection. LP and RD contributed to data curation. LP, GMQ-G, LB, RH, and CL contributed to data analysis. LB, RH, DL, CL, JS, and MV contributed to data interpretation. LB, CL, and MV contributed to the figure design. GMQ-G, JS, AAS, VS, and RD contributed to expert review. LP, GMQ-G, LB, DL, CL, MV, PM, SH, and RD contributed to manuscript preparation and review.

Declaration of interests

LP and VS serve on the medical advisory board of Ibex and report personal fees from Ibex, outside of the submitted work. LB and CL are authors on pending patents US 62/743,559 and US 62/981,925 (including System & Methods for Personalization and Optimization of Digital Pathology Analysis and System and Method of Managing Workflow of Examination of Pathology Slides). AAS reports personal fees for expert consultancy from Ibex, outside of the submitted work. All other authors declare no competing interests.

Data sharing

The software used in the current study is Ibex proprietary IP and cannot be shared. The data collected during this study is patient data obtained under Ethical Committees approval and cannot be shared. De-identified raw whole slide imaging data described in this study are available for viewing online and for download upon request.

Acknowledgments

This study was funded by Ibex Medical Analytics.

References

- 1 Rawla P. Epidemiology of prostate cancer. *World J Oncol* 2019; **10**: 63–89.
- 2 Matoso A, Epstein JI. Defining clinically significant prostate cancer on the basis of pathological findings. *Histopathology* 2019; **74**: 135–45.
- 3 Metter DM, Colgan TJ, Leung ST, Timmons CF, Park JY. Trends in the US and Canadian pathologist workforces from 2007 to 2017. *JAMA Netw Open* 2019; **2**: e194337.
- 4 Evans AJ, Bauer TW, Bui MM, et al. US Food and Drug Administration approval of whole slide imaging for primary diagnosis: a key milestone is reached and new questions are raised. *Arch Pathol Lab Med* 2018; **142**: 1383–87.
- 5 Hipp J, Monaco J, Kunju LP, et al. Integration of architectural and cytologic driven image algorithms for prostate adenocarcinoma identification. *Anal Cell Pathol (Amst)* 2012; **35**: 251–65.
- 6 Campanella G, Hanna MG, Geneslaw L, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med* 2019; **25**: 1301–09.
- 7 Nir G, Karimi D, Goldenberg SL, et al. Comparison of artificial intelligence techniques to evaluate performance of a classifier for automatic grading of prostate cancer from digitized histopathologic images. *JAMA Netw Open* 2019; **2**: e190442.
- 8 Bulten W, Pinckaers H, van Boven H, et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *Lancet Oncol* 2020; **21**: 233–41.
- 9 Rice-Stitt T, Valencia-Guerrero A, Cornejo KM, Wu CL. Updates in histologic grading of urologic neoplasms. *Arch Pathol Lab Med* 2020; **144**: 335–43.
- 10 Fraggetta F. Clinical-grade computational pathology: alea iacta est. *J Pathol Inform* 2019; **10**: 38.
- 11 Yang C, Humphrey PA. False-negative histopathologic diagnosis of prostatic adenocarcinoma. *Arch Pathol Lab Med* 2020; **144**: 326–34.
- 12 Steinberg DM, Sauvageot J, Piantadosi S, Epstein JI. Correlation of prostate needle biopsy and radical prostatectomy Gleason grade in academic and community settings. *Am J Surg Pathol* 1997; **21**: 566–76.
- 13 Kronz JD, Milord R, Wilentz R, Weir EG, Schreiner SR, Epstein JI. Lesions missed on prostate biopsies in cases sent in for consultation. *Prostate* 2003; **54**: 310–14.
- 14 Truesdale MD, Cheatham PJ, Turk AT, et al. Gleason score concordance on biopsy-confirmed prostate cancer: is pathological re-evaluation necessary prior to radical prostatectomy? *BJU Int* 2011; **107**: 749–54.
- 15 Goodman M, Ward KC, Osunkoya AO, et al. Frequency and determinants of disagreement and error in Gleason scores: a population-based study of prostate cancer. *Prostate* 2012; **72**: 1389–98.
- 16 Raciti P, Sue J, Ceballos R, et al. Novel artificial intelligence system increases the detection of prostate cancer in whole slide images of core needle biopsies. *Mod Pathol* 2020; published online May 11. <https://doi.org/10.1038/s41379-020-0551-y>.
- 17 Laifenfeld D, Sandbank J, Linhart C, et al. Performance of an AI-based cancer diagnosis system in France's largest network of pathology institutes. 31st European Congress of Pathology; Sept 7–11, 2019; Nice, France (abstr 043).
- 18 Owens SR, Wiehagen LT, Kelly SM, et al. Initial experience with a novel pre-sign-out quality assurance tool for review of random surgical pathology diagnoses in a subspecialty-based university practice. *Am J Surg Pathol* 2010; **34**: 1319–23.
- 19 Parwani AV. Automated diagnosis and Gleason grading of prostate cancer—are artificial intelligence systems ready for prime time? *J Pathol Inform* 2019; **10**: 41.
- 20 Goldenberg SL, Nir G, Salcudean SE. A new era: artificial intelligence and machine learning in prostate cancer. *Nat Rev Urol* 2019; **16**: 391–403.
- 21 Ström P, Kartasalo K, Olsson H, et al. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. *Lancet Oncol* 2020; **21**: 222–32.
- 22 Arvaniti E, Fricker KS, Moret M, et al. Automated Gleason grading of prostate cancer tissue microarrays via deep learning. *Sci Rep* 2018; **8**: 12054.
- 23 Nagpal K, Foote D, Liu Y, et al. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. *NPJ Digit Med* 2019; **2**: 48.
- 24 Nir G, Hor S, Karimi D, et al. Automatic grading of prostate cancer in digitized histopathology images: learning from multiple experts. *Med Image Anal* 2018; **50**: 167–80.
- 25 Lucas M, Jansen I, Savci-Heijink CD, et al. Deep learning for automatic Gleason pattern classification for grade group determination of prostate biopsies. *Virchows Arch* 2019; **475**: 77–83.

For de-identified raw whole slide imaging data see <http://image.upmc.edu:8080/ibex/view.apml>

- 26 Sparks R, Madabhushi A. Statistical Shape Model for Manifold Regularization: Gleason grading of prostate histology. *Comput Vis Image Underst* 2013; **117**: 1138–46.
- 27 Hossain A, Arimura H, Kinoshita F, et al. Automated approach for estimation of grade groups for prostate cancer based on histological image feature analysis. *Prostate* 2020; **80**: 291–302.
- 28 Harnden P, Shelley MD, Clements H, et al. The prognostic significance of perineural invasion in prostatic cancer biopsies: a systematic review. *Cancer* 2007; **109**: 13–24.
- 29 Paner GP, Gandhi J, Choy B, Amin MB. Essential updates in grading, morphotyping, reporting, and staging of prostate carcinoma for general surgical pathologists. *Arch Pathol Lab Med* 2019; **143**: 550–64.
- 30 Leo P, Lee G, Shih NN, Elliott R, Feldman MD, Madabhushi A. Evaluating stability of histomorphometric features across scanner and staining variations: prostate cancer diagnosis from whole slide images. *J Med Imaging (Bellingham)* 2016; **3**: 047502.

Pantanowitz, Liron

From: Permissions Helpdesk <permissionshelpdesk@elsevier.com>
Sent: Thursday, September 3, 2020 3:18 PM
To: Pantanowitz, Liron
Subject: Re: Request to use my article for my PhD thesis [200901-012026]

External Email - Use Caution



Dear Liron Pantanowitz,

Thank you for your query.

Please note that, as one of the authors of this article, you retain the right to reuse it in your thesis/dissertation. You do not require formal permission to do so. You are permitted to post this Elsevier article online if it is embedded within your thesis. You are also permitted to post your Author Accepted Manuscript online.

However posting of the final published article is prohibited.

"As per our Sharing Policy, authors are permitted to post the Accepted version of their article on their institutional repository – as long as it is for internal institutional use only.

It can only be shared publicly on that site once the journal-specific embargo period has lapsed. For a list of embargo periods please see: Embargo List.

You are not permitted to post the Published Journal Article (PJA) on the repository."

Please feel free to contact me if you have any queries.

Regards,

Kaveri

Permissions Helpdesk
ELSEVIER | Operations

From: Administrator
Date: 01/09/2020 01.06 PM

CHAPTER 5: PUBLISHED ARTICLE 4

Pantanowitz L, Michelow P, Hazelhurst S, Kalra S, Choi C, Shah S, Babaie M, Tizhoosh HR. A digital pathology solution to resolve the tissue floater conundrum. *Archives of Pathology & Laboratory Medicine*, 2021; 145(3):359-364.
<https://pubmed.ncbi.nlm.nih.gov/32886759/>

Abstract

Context: Pathologists may encounter extraneous pieces of tissue (tissue floaters) on glass slides because of specimen cross-contamination. Troubleshooting this problem, including performing molecular tests for tissue identification if available, is time consuming and often does not satisfactorily resolve the problem.

Objective: To demonstrate the feasibility of using an image search tool to resolve the tissue floater conundrum.

Design: A glass slide was produced containing 2 separate hematoxylin and eosin (H&E)-stained tissue floaters. This fabricated slide was digitized along with the 2 slides containing the original tumors used to create these floaters. These slides were then embedded into a dataset of 2325 whole slide images comprising a wide variety of H&E stained diagnostic entities. Digital slides were broken up into patches and the patch features converted into barcodes for indexing and easy retrieval. A deep learning-based image search tool was employed to extract features from patches via barcodes, hence enabling image matching to each tissue floater.

Results: There was a very high likelihood of finding a correct tumor match for the queried tissue floater when searching the digital database. Search results repeatedly yielded a correct match within the top 3 retrieved images. The retrieval accuracy

improved when greater proportions of the floater were selected. The time to run a search was completed within several milliseconds.

Conclusions: Using an image search tool offers pathologists an additional method to rapidly resolve the tissue floater conundrum, especially for those laboratories that have transitioned to going fully digital for primary diagnosis.

A Digital Pathology Solution to Resolve the Tissue Floater Conundrum

Liron Pantanowitz, MD; Pamela Michelow, MBBCh; Scott Hazelhurst, PhD; Shivam Kalra, BSc; Charles Choi, BSc; Sultaan Shah, BSc; Morteza Babaie, PhD; Hamid R. Tizhoosh, PhD

• **Context.**—Pathologists may encounter extraneous pieces of tissue (tissue floaters) on glass slides because of specimen cross-contamination. Troubleshooting this problem, including performing molecular tests for tissue identification if available, is time consuming and often does not satisfactorily resolve the problem.

Objective.—To demonstrate the feasibility of using an image search tool to resolve the tissue floater conundrum.

Design.—A glass slide was produced containing 2 separate hematoxylin and eosin (H&E)-stained tissue floaters. This fabricated slide was digitized along with the 2 slides containing the original tumors used to create these floaters. These slides were then embedded into a dataset of 2325 whole slide images comprising a wide variety of H&E stained diagnostic entities. Digital slides were broken up into patches and the patch features converted into barcodes for indexing and easy retrieval. A deep learn-

ing-based image search tool was employed to extract features from patches via barcodes, hence enabling image matching to each tissue floater.

Results.—There was a very high likelihood of finding a correct tumor match for the queried tissue floater when searching the digital database. Search results repeatedly yielded a correct match within the top 3 retrieved images. The retrieval accuracy improved when greater proportions of the floater were selected. The time to run a search was completed within several milliseconds.

Conclusions.—Using an image search tool offers pathologists an additional method to rapidly resolve the tissue floater conundrum, especially for those laboratories that have transitioned to going fully digital for primary diagnosis.

(*Arch Pathol Lab Med.* 2021;145:359–364; doi: 10.5858/arpa.2020-0034-OA)

In routine clinical practice, pathologists may encounter extraneous pieces of tissue on glass slides that could be because of contamination from other specimens. These are

typically called “tissue floaters.” A prior review analyzing this quality issue in 275 laboratories reported approximately 3% of their pathology slides could potentially contain such tissue contaminants.¹ Tissue contamination can occur at various steps in anatomic pathology, including grossing or downstream in the histology laboratory during tissue processing, slide preparation, or with staining. Common causes of this problem appear to be because of the pickup of floaters in the water bath during sectioning,² or during staining of slides when deparaffinized tissue can easily fragment.³ The dilemma pathologists often face is whether such a tissue floater truly belongs to the case in question, or if instead it represents a true contaminant from another patient’s sample in which case it should be ignored. This poses a major conundrum when the tissue floater contains malignant cells, and is unlikely to be derived from an obviously different organ or tumor.

There are currently several measures a pathologist can employ to troubleshoot the tissue floater problem. They could ask their colleagues on clinical service if they have recently encountered a case that morphologically resembles the tissue floater (eg, by emailing them a photo of the tissue floater). They could also physically inspect the tissue block (if available) and/or order a recut section to see if this subsequent slide harbors this floater in the same position on the slide, which may help pinpoint where potential cross-contamination occurred. Akin to forensic analysis, some laboratories have implemented molecular techniques (eg, DNA fingerprinting for tissue identity testing) to try resolve

Accepted for publication May 11, 2020.
Published online July 15, 2020.

From the Department of Pathology, University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania (Pantanowitz); Department of Anatomical Pathology, University of the Witwatersrand and National Health Laboratory Services, Johannesburg, South Africa (Pantanowitz, Michelow); School of Electrical & Information Engineering and Sydney Brenner Institute for Molecular Bioscience, University of the Witwatersrand, Johannesburg, South Africa (Hazelhurst); Kimia Lab, University of Waterloo, Waterloo, Ontario, Canada (Kalra, Babaie, Tizhoosh); and Huron Digital Pathology, Engineering Department, St. Jacobs, Ontario, Canada (Kalra, Choi, Shah). Pantanowitz is now with the Department of Pathology & Clinical Labs, University of Michigan, Ann Arbor.

Huron Digital Pathology provided financial support for this project, the Ontario government provided ORF-RE funds, and the federal Canadian government provided a NSERC-CRD Grant to support this project.

Pantanowitz is on the medical advisory board for Leica, Ibex, and Hologic and consults for Hamamatsu. Tizhoosh is on the advisory board for Huron Digital Pathology. Kalra has an industrial internship at Huron Digital Pathology. The other authors have no relevant financial interest in the products or companies described in this article.

Corresponding author: Liron Pantanowitz, MD, Department of Pathology & Clinical Labs, University of Michigan, NCRC Bldg 35, Room 36-1221-35, 2800 Plymouth Rd, Ann Arbor, MI 48109-2800 (email: lironp@med.umich.edu).

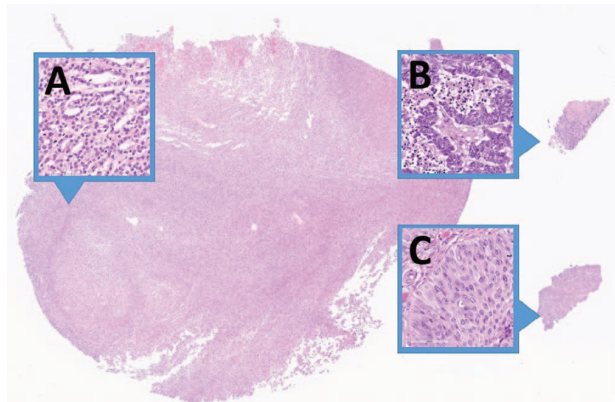


Figure 1. Fabricated slide containing a section of renal cell carcinoma (A) and 2 adjacent separate colon cancer (B) and bladder cancer (C) tissue floaters (hematoxylin and eosin stain, insets shown at $\times 20$ magnification).

this problem by dissecting, testing, and then comparing the molecular results of the tissue floater to the adjacent patient sample on the glass slide.⁴⁻⁸

With increasing global adoption of whole slide imaging (WSI) in pathology departments, it is now feasible to exploit digital pathology to resolve the tissue floater conundrum. WSI refers to the digitization of glass slides, using a slide scanner, in order to generate corresponding digital slides that can be remotely viewed on a computer monitor.^{9,10} Several pathology laboratories have already transitioned to prospectively digitizing most of their slides in order for pathologists to render a primary diagnoses, without having to examine the original glass slides.^{11,12} In such a digital environment, the proposed digital solution when a tissue floater is encountered would be to annotate (ie, select) the suspected tissue floater region of interest in the image, and to then search their archived database of recently scanned slides for a matching digital slide that is the probable source of this specimen processing error. Indeed, several publications have shown researchers have been able to develop sophisticated machine and/or deep learning algorithms that can suitably classify histopathological images and perform content-based image retrieval (CBIR).¹³⁻²⁰ CBIR is the application of computer vision methods to search for digital images in a large database.^{21,22}

The aim of this study was to demonstrate the feasibility of using a dedicated histopathology image search tool in order to resolve the tissue floater conundrum. This article reports the success of such an image search algorithm developed to match and retrieve similar digital slides or regions of interest using pixels.

MATERIALS AND METHODS

Institutional review board approval was obtained for this study (University of Pittsburgh, Pennsylvania, STUDY18100084, PRO17120392; University of Witwatersrand, Johannesburg, South Africa clearance certificate M191003).

Tissue Floater Slide Creation

For the purposes of this experiment, 3 glass slides were prepared using freshly discarded adult human tumor tissue (Figure 1). Fabricated tissue floater slides were artificially created that contained (1) a large central portion of renal cell carcinoma obtained from a type 1 (basophilic) papillary renal cell carcinoma,

Table 1. Top 20 Primary Sites With the Highest Number of Whole Slide Imaging (WSI) in the Dataset

Primary Site	Number of WSIs
Brain	344
Breast	198
Lung	213
Kidney	192
Uterus	133
Colorectal	123
Skin	124
Soft tissue	121
Head and neck	100
Thyroid	99
Prostate	94
Stomach	69
Liver	65
Adrenal gland	65
Cervix	51
Bladder	43
Testis	34
Esophagus	33
Pancreas	20
Ovary	19

and (2) placed toward the edge of the slide 2 separate additional smaller portions of tissue ("tissue floaters") obtained from a moderately differentiated colon adenocarcinoma and a urinary bladder high-grade papillary urothelial carcinoma. To avoid having these slides stand out as extraordinary in the study the pathology cases selected were relatively easy to diagnose and typical of what would be encountered in routine practice. Also, the prepared slides were stained with hematoxylin and eosin (H&E) according to routine staining protocols. All 3 slides were then entirely digitized at $\times 40$ magnification using an Aperio AT2 whole slide scanner (Leica Biosystems). The quality of these digital slides was checked to avoid inclusion of unique identifiers and/or artifacts.

Pathology Digital Slide Datasets

The aforementioned WSIs were embedded into 2 datasets of digital slides. The first dataset was established by randomly selecting 300 de-identified WSIs (.svs file format) of H&E-stained surgical pathology cases from the teaching files at the University of Pittsburgh Medical Center, Pittsburgh, PA ("UPMC dataset"). These archival slides were scanned at $\times 40$ magnification using an Aperio ScanScope XT instrument (Leica Biosystems). These WSIs included cases from a wide variety of anatomic sites (eg, colon, brain, thyroid, prostate, breast, kidney, salivary gland, skin, soft tissue, etc) exhibiting varied diagnostic pathologic entities (ie, reactive, inflammatory, benign neoplasms, and malignancies). The other dataset employed in this study was obtained from the publicly available digital pathology slide archive offered by The Cancer Genome Atlas (TCGA) program (<https://portal.gdc.cancer.gov>). A total of 2025 WSIs were randomly selected and downloaded from the TCGA ("TCGA dataset"). An average WSI was approximately $45\,000 \times 45\,000$ pixels. Digital slides of low quality (eg, very poor staining, low resolution, large regions out of focus) were eliminated; magnifications less than $\times 20$ and blurry patches understandably prevent the image search to perform well and poor staining negatively affects extraction of deep features. The TCGA dataset incorporated at least 33 different diagnostic entities from 25 anatomic locations. TCGA slides of frozen sections and those slides with manual annotations (pen markings) present were included. All digital slides were labeled with both the type of malignancy (primary diagnosis) and the affected organ (primary site). This label was

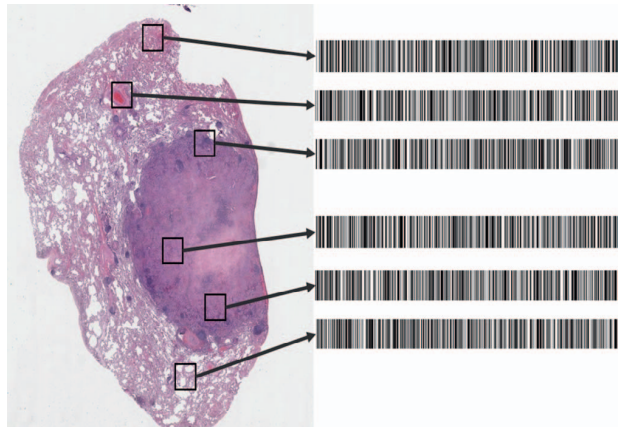


Figure 2. Schematic illustration of the general idea of using barcodes for image representation: whole slide image indexed by converting separate patches into barcodes.

assigned to the entire WSI and no individual region was delineated. Table 1 shows the top 20 primary sites with the highest number of WSIs in the combined dataset (ie, UPMC + TCGA datasets).

Computing Platform

All experiments were performed on a Dell EdgeServer Ra with 2x Intel(R) Xeon(R) Gold 5118 (12 cores, 2.30 GHz), 4x Telsa V100 (v-RAM 32 GB each; only 2 graphics processing units [GPUs] were used), and 394-GB random access memory. The code for indexing was written in C/C++. The user interface components were written in multiple languages, but mostly in Python and JavaScript. The high-end GPU power was necessary for indexing the large archives, which was a 1-time task for existing repositories. For daily usage of image search ordinary (low-cost) central processing unit/GPU power will suffice as the barcodes (see below) enable efficient search in large archives.

Image Search Tool

Through an ensemble approach (using a cohort of different algorithms) a reliable search engine prototype was developed that exploited the strengths of both supervised (trained deep networks) and unsupervised (clustering and search) computational methods for image processing.²³ This image search tool thus included segmentation and clustering algorithms, deep networks, and distance metrics for search and retrieval. Whereas deep networks are supervised methods and require extensive training with labeled data, the search itself is unsupervised with no prior training. Using a pretrained deep network without fine-tuning does not constitute direct supervision because it is used for feature extraction without any adjustment. This allows our approach to be independent of manually delineated WSIs. We used DenseNet-121, which is publicly available. Image segmentation was used to distinguish tissue from white background. WSIs were broken into patches or tiles at fixed sizes (eg, $500 \times 500 \mu\text{m}^2$ at $\times 20$) with no overlap. The patches were grouped into categories via clustering methods (eg, k-means algorithm) and passed through pretrained artificial neural networks for feature mining. Each feature vector was converted into a linear barcode (Figure 2) and this “bunch of barcodes” indexing process was used to accelerate the search retrieval process.²⁴ The barcode generation only contained binarization of gradient change of deep features. Multiple similarity measures were examined to further increase the matching rate when comparing images. Retrieved image patches were ranked from most to least likelihood of being similar to the queried image (Figure 3) and these results were displayed in a gallery format for the end user to review and interpret. Rank (defined as the rating of a suitable match) was determined by the similarity of the suspected patch with all other patches in the archive, measured through distance

Table 2. Initial Matched Results for Each Tissue Floater (UPMC 300 Whole Slide Imaging Dataset)

Selected Match	For Bladder Tumor Floater	For Colon Tumor Floater
First match is correct	80%	66%
Second match is correct	87%	73%
Third match is correct	93%	73%
First and second match are correct	80%	90%
Top 3 matches are correct	80%	90%

calculation between their corresponding barcodes. The more similar (ie, best ranked result) the patch is the smaller is the difference between barcodes. All matched patches were sorted based on this difference (least different is ranked first and so on).

Any tissue fragment can potentially be selected by the end user (pathologist) for searching the archive. As such, a search is manually triggered by the pathologist. The smallest patch size that can be indexed and searched is $500 \times 500 \mu\text{m}$ ($\sim 1000 \times 1000$ pixels at $\times 20$); any floater smaller than that may not be detectable.

Search Tool Evaluation

After the aforementioned 3 fabricated slides were scanned, indexed, and mixed among millions of image patches from the 2 datasets, the number of patches was reduced to approximately 16 000 patches through clustering (empirically set to 9 groups), whereas only 5% of each cluster was selected to represent a WSI. The search tool was then used to try and identify the matching slide that belonged to each tissue floater (Figure 4). The search was conducted using variable percentages of floater sampling (ie, 5%–100% of tissue floater region selected). The detection accuracy was measured by running each sample 100 times (manually and by automation) and calculating the median rank of a correct detection among search results, as well as the best and the worst rank of the detected floater among the search results using a 95% CI.

RESULTS

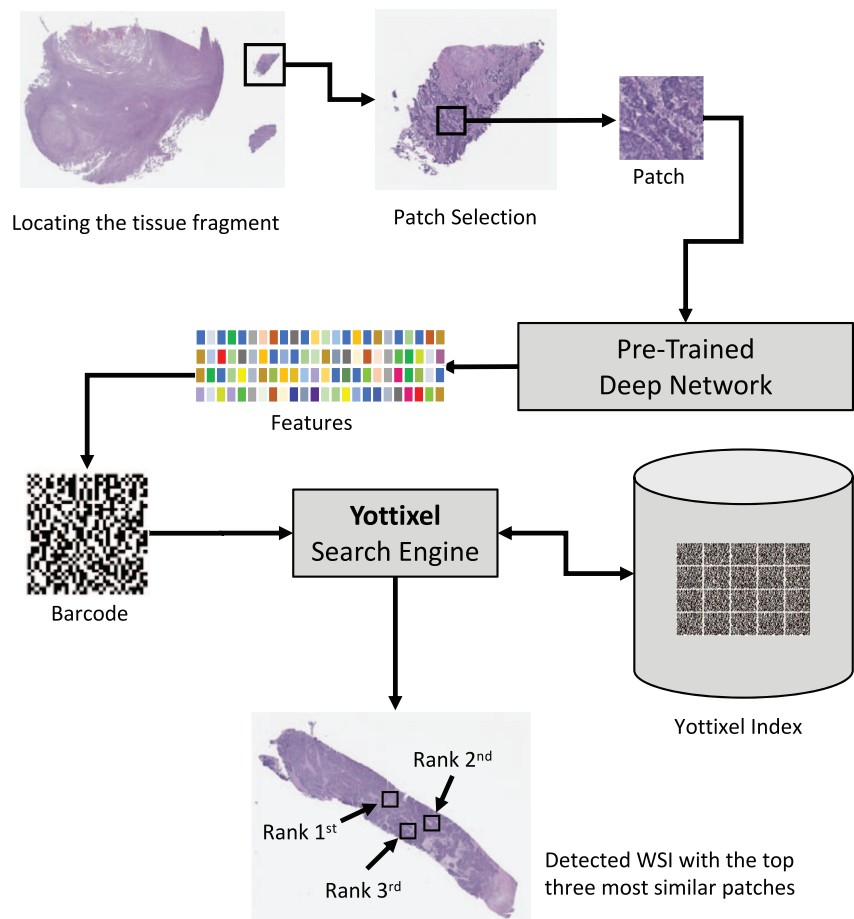
Initial Algorithm Validation

The image search platform was qualitatively tested using the 300 UPMC WSIs. For this pilot experiment, 100 sample regions were randomly selected in order to demonstrate the search tool’s capability. The retrieval results were evaluated by a pathologist (LP) and converted into an accuracy value. Table 2 shows the accuracy in predicting the correct query image using only the top 3 first search results (ie, the best 3 matches).

Image Search Results

The image search results for matching tissue floaters when using the UPMC 300 WSI pilot dataset showed that the median rank best result for both the bladder and colon tumor was 1 (95% CI = 1) when selecting 5% up to 100% of the floater region. Table 3 shows search results when trying to match images to tissue floaters among all 2325 digital slides (ie, 300 UPMC + 2025 National Cancer Institute). Table 4 shows the retrieval accuracy for matching image patches to both tissue floaters when a manual search was performed. The retrieval accuracy (ie, successful rank) of finding a correct match when undertaking a search greatly increased when more of the tissue floater region was selected. Searches were computationally the most expensive to perform when selecting 100% of the tissue floater. The time to perform a search measured was very short and never more than several milliseconds.

Figure 3. Schematic diagram showing how the origin of a suspected floater gets detected. The process starts with locating the suspicious tissue fragment. A selected patch from the fragment is then fed into a pretrained deep network to extract features. The search engine then receives a generated barcode to search within the “Yottixel Index” that contains barcodes of many patches of many whole slide images (WSIs). Finally, the origin of the floater is recognized by investigating the top ranked patches.



DISCUSSION

Tissue floaters because of cross-contamination between specimens are a noteworthy cause of diagnostic error in anatomic pathology.² Such mishaps can potentially lead to a serious misdiagnosis. Therefore, apart from implementing a quality assurance program with process improvement in order to mitigate this diagnostic error it is also important that pathologists can adequately resolve this dilemma when encountered in clinical practice. If a tissue block contains the misidentified piece, deeper levels may help verify tissue contamination occurred at the grossing bench or when embedding the paraffin block. However, performing a root cause analysis that may involve ordering recuts is time consuming and often does not satisfactorily resolve the problem. Moreover, molecular testing for tissue identification may not be readily available for many pathology laboratories, and the results could be missed as DNA may be altered by staining or other tissue processing steps. The image search tool demonstrated in our study offers

pathologists an additional method to resolve the tissue floater conundrum, especially for those laboratories that have transitioned to using WSI for routine diagnostic work.

The accuracy of the search tool to retrieve matching image patches in our study improved with each iteration of the algorithm. The accuracy when searching for a match to the bladder tumor floater improved from 80% to 93% initially up to 95% and for the colon tissue floater it improved from 66% to 73% to 90% when performing manual searches. After the algorithm was fully developed, we witnessed a near perfect match for top-ranked retrieved image patches. These data further show that when a greater proportion of the tissue floater is selected the search is more apt to almost always yield a useful match with just the first ranked retrieved image. With adequate computing capability the time to run a search was only several milliseconds, which is far quicker than the alternate methods (eg, DNA fingerprinting) proposed for troubleshooting a slide that includes a tissue floater. Once a digital pathology system with this

Table 3. Image Search Results That Match Tissue Floaters (UPMC + NCI 2325 Whole Slide Imaging Dataset)

Tissue Floater	5% of Floater	10% of Floater	25% of Floater	50% of Floater	75% of Floater	100% of Floater
Bladder tumor: median rank of best result [95% CI]	1 [1, 1.25]	1 [1, 1]	1 [1, 1]	1 [1, 1]	1 [1, 1]	1 [1, 1]
Colon tumor: median rank of best result [95% CI]	1 [3, 7]	1 [2, 3]	1 [1, 2]	1 [1, 2]	1 [1, 1]	1 [1, 1]

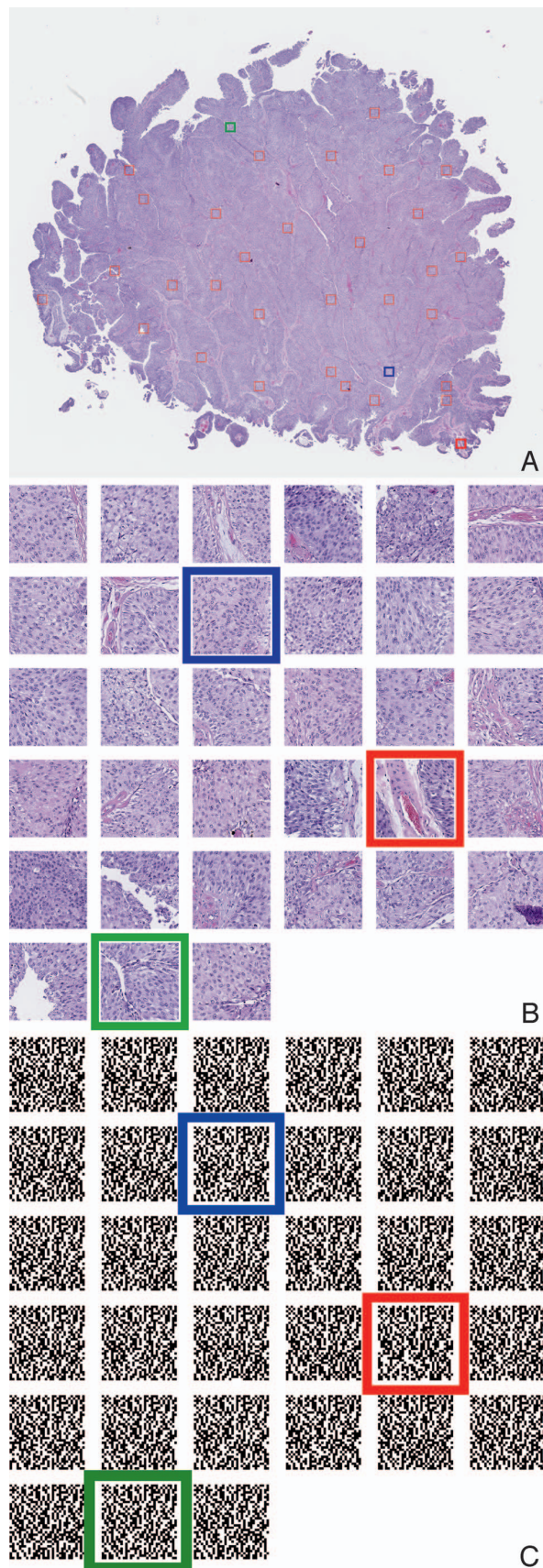


Table 4. Top 5 Retrieved Patches for Manual Search

Position Within Search Results	Accuracy for Bladder WSI	Accuracy for Colon WSI
The first retrieved patch	90%	90%
The second retrieved patch	95%	75%
The third retrieved patch	85%	85%
The fourth retrieved patch	95%	70%
The fifth retrieved patch	95%	70%

Abbreviation: WSI, whole slide imaging.

proposed artificial intelligence tool is established it may also be cheaper to workup tissue floaters this way rather than with molecular testing. It is unclear if the superior retrieval results noted with the bladder tumor search compared with the colon tumor floater were because of the fact that the database contained fewer overall bladder cases ($n = 43$) than colorectal digital slides ($n = 123$). To the best of our knowledge, we are not aware of a similar published study using image search to resolve tissue floaters on pathology slides.

The digital pathology solution used in this study applied image classification and CBIR (or query by image content) technology. Such technology relies on a variety of techniques including statistics, pattern recognition, traditional computer vision, and deep learning algorithms.^{21,25–27} With CBIR the search tool uses the contents (eg, features such as color, shape, texture, spatial relationship, etc) within each pixel of the image rather than using annotations or metadata (eg, tags or diagnostic labels associated with the images). This accordingly allows relevant images to be automatically retrieved from a large database that match the contents of the queried image. For very large databases it is impractical to have humans manually annotate each image. To address this challenge, we employed unsupervised learning and deep neural networks when developing the image search tool. We also minimized the demand for expensive computer equipment when running a search query on large WSI files by converting individual image patches into barcodes. The barcodes allowed the WSIs in this study to be completely indexed, precisely recognized, and easily retrieved.

Google's well-known image search tool, called Reverse Image Search (<https://images.google.com>), allows users to search websites with a picture for similar images. This search tool can also be used to decipher medical images.²⁸ Hanna and Pantanowitz²⁹ previously tested Google's freely available Reverse Image Search tool for diagnostic surgical pathology. They found when employing this online tool to find related images on the web using histopathology images alone as input their search results were unreliable to make a histopathology diagnosis. However, after adding the organ and term "cancer" to their image search they greatly increased the diagnostic specificity of their search results (eg, up to 83% for certain neoplasms). Researchers from Google have since refined their image search capability and developed a deep learning-based reverse image search tool

Figure 4. (A) Indexing of a sample whole slide imaging (scan with bladder tumor) (B) yielding 33 patches to build a mosaic. (C) Corresponding barcodes of the mosaic can be generated using a MinMax algorithm. The 3 barcodes that match the highlighted patches are shown.

specifically for histopathology images called Similar Image Search for Histopathology (SMILY).³⁰ The authors of this article demonstrated that SMILY was able to retrieve images with similar organ site, histologic features, and cancer grade compared with the original query image. SMILY has great potential for diagnosis, research, and education. If it becomes freely available, perhaps it can also be validated for working-up cases with tissue floaters.

As pathology departments continually scan their glass slides, they will begin to amass large digital databases. If these WSIs are appropriately stored, indexed, and linked with corresponding metadata (eg, pathology reports, annotations, patient outcomes) these archives will become increasingly valuable in the emerging era of artificial intelligence.^{31,32} This will allow pathology laboratories to not only exploit artificial intelligence-based image search, but additionally leverage their “Big Data” for other computational pathology applications.^{31,33} Further work is underway to clinically validate our image search tool to deal with actual tissue floaters in a clinical work environment and to enhance the user interface to make it easy for pathologists to employ in routine practice.

The results in this study are partly based upon WSI offered by the TCGA Research Network (<https://www.cancer.gov/tcga>). We thank the company Huron Digital Pathology for their financial support of this project, the Ontario government for ORF-RE funds, and the federal Canadian government for an NSERC-CRD Grant to support this project.

References

- Gephardt GN, Zarbo RJ. Extraneous tissue in surgical pathology: a College of American Pathologists Q-probes study of 275 laboratories. *Arch Pathol Lab Med*. 1996;120(11):1009–1014.
- Layfield LJ, Witt BL, Metzger KG, Anderson GM. Extraneous tissue: a potential source for diagnostic error in surgical pathology. *Am J Clin Pathol*. 2011;136(5):767–772.
- Platt E, Sommer P, McDonald L, Bennett A, Hunt J. Tissue floaters and contaminants in the histology laboratory. *Arch Pathol Lab Med*. 2009;133(6):973–997.
- Hunt JL, Swalsky P, Sasatomi E, Niehouse L, Bakker A, Finkelstein SD. A microdissection and molecular genotyping assay to confirm the identity of tissue floaters in paraffin-embedded tissue blocks. *Arch Pathol Lab Med*. 2003;127(2):213–217.
- Demetrick D, Hay R, Wahgray R. DNA “fingerprint” evaluation of clinical specimen contamination: application to cytopathology. *Pathol Case Rev*. 2006;11(2):98–102.
- Venditti M, Hay RW, Kulaga A, Demetrick DJ. Diagnosis of ectopic tissue versus contamination by genetic fingerprinting in a routine surgical pathology specimen. *Hum Pathol*. 2007;38(2):378–382.
- Mosse CA, Stumph JR, Best DH, Vnencak-Jones CL. A B-cell lymphoma diagnosed in “floater” tissue: implications of the diagnosis and resolution of a laboratory error. *Am J Med Sci*. 2009;338(3):248–251.
- Bossuyt V, Buza N, Ngo NT, et al. Cancerous ‘floater’: a lesson learned about tissue identity testing, endometrial cancer and microsatellite instability. *Mod Pathol*. 2013;26(9):1264–1269.
- Mukhopadhyay S, Feldman MD, Abels E, et al. Whole slide imaging versus microscopy for primary diagnosis in surgical pathology: a multicenter blinded randomized noninferiority study of 1992 cases (pivotal study). *Am J Surg Pathol*. 2018;42(1):39–52.
- Borowsky AD, Glassy EF, Wallace WD, et al. Digital whole slide imaging compared with light microscopy for primary diagnosis in surgical pathology: a multicenter, double-blinded, randomized study of 2045 cases [published online February 14, 2020]. *Arch Pathol Lab Med*. doi: 10.5858/arpa.2019-0569-OA.
- Stathonikos N, Veta M, Huisman A, van Diest PJ. Going fully digital: perspective of a Dutch academic pathology lab. *J Pathol Inform*. 2013;4:15.
- Fragetta F, Garazzo S, Zannoni GF, Pantanowitz L, Rossi ED. Routine digital pathology workflow: the Catania experience. *J Pathol Inform*. 2017;8:51.
- Caicedo JC, González FA, Romero E. Content-based histopathology image retrieval using a kernel-based semantic annotation framework. *J Biomed Inform*. 2011;44(4):519–528.
- Qi X, Wang D, Rodero I, et al. Content-based histopathology image retrieval using CometCloud. *BMC Bioinformatics*. 2014;15:287.
- Sridhar A, Doyle S, Madabhushi A. Content-based image retrieval of digitized histopathology in boosted spectrally embedded spaces. *J Pathol Inform*. 2015;6:41.
- Kwak JT, Hewitt SM, Kajdacsy-Balla AA, Sinha S, Bhargava R. Automated prostate tissue referencing for cancer detection and diagnosis. *BMC Bioinformatics*. 2016;17(1):227.
- Zhu S, Li Y, Kalra S, Tizhoosh HR. Multiple disjoint dictionaries for representation of histopathology images. *J Vis Commun Image Represent*. 2018;55:243–252.
- Ma Y, Jiang Z, Zhang H, et al. Generating region proposals for histopathological whole slide image retrieval. *Comput Methods Programs Biomed*. 2018;159:1–10.
- Zheng Y, Jiang Z, Zhang H, et al. Histopathological whole slide image analysis using context-based CBIR. *IEEE Trans Med Imaging*. 2018;37(7):1641–1652.
- Zheng Y, Jiang Z, Zhang H, et al. Size-scalable content-based histopathological image retrieval from database that consists of WSIs. *IEEE J Biomed Health Inform*. 2018;22(4):1278–1287.
- Latif A, Rasheed A, Sajid U, et al. Content-based image retrieval and feature extraction: a comprehensive review. *Mathematical Problems in Engineering*. Hindawi website. <https://www.hindawi.com/journals/mpe/2019/9658350/>. Accessed December 25, 2019.
- Akgül CB, Rubin DL, Napel S, Beaulieu CF, Greenspan H, Acar B. Content-based image retrieval in radiology: current status and future directions. *J Digit Imaging*. 2011;24(2):208–222.
- Babaie M, Kalra S, Sriram A, et al. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2017, pp. 8–16. http://openaccess.thecvf.com/content_cvpr_2017_workshops/w8/papers/Babaie_Classification_and_Retrieval_CVPR_2017_paper.pdf. Accessed December 25, 2019.
- Kalra S, Choi C, Shah S, Pantanowitz L, Tizhoosh HR. Yottixel - an image search engine for large archives of histopathology whole slide images. 2019. <https://arxiv.org/abs/1911.08748>. Accessed December 25, 2019.
- Wang XY, Zhang BB, Yang HY. Content-based image retrieval by integrating color and texture features. *Multimedia Tools and Applications*. 2014;68(3):545–569.
- Sklan JE, Plassard AJ, Fabbri D, Landman BA. Toward content based image retrieval with deep convolutional neural networks. *Proc SPIE Int Soc Opt Eng*. 2015;9417.
- Zhou W, Li H, Tian Q. Recent advance in content-based image retrieval: a literature survey. 2017. <https://arxiv.org/abs/1706.06064>. Accessed December 25, 2019.
- Mamrosh JL, Moore DD. Using Google reverse image search to decipher biological images. *Curr Protoc Mol Biol*. 2015;111:19.13.1–19.13.4.
- Hanna M, Pantanowitz L. Utility of Google image search for diagnostic surgical pathology. *J Pathol Inform*. 2017;8:26.
- Hegde N, Hipp JD, Liu Y, et al. Similar image search for histopathology: SMILY. *NPJ Digit Med*. 2019;2:56.
- Tizhoosh HR, Pantanowitz L. Artificial intelligence and digital pathology: challenges and opportunities. *J Pathol Inform*. 2018;9:38.
- Niazi MKK, Parwani AV, Gurcan MN. Digital pathology and artificial intelligence. *Lancet Oncol*. 2019;20:e253–e261.
- Tizhoosh H. Searching is intelligence. Image retrieval – the next revolution in pathology. *The Pathologist*. <https://thepathologist.com/diagnostics/searching-is-intelligence>. Accessed December 25, 2019.

Pantanowitz, Liron

From: Hilary Price (s) <hprice@cap.org>
Sent: Thursday, September 3, 2020 10:09 AM
To: Pantanowitz, Liron
Subject: RE: Permission

External Email - Use Caution

Hi Dr. Pantanowitz,

The permission letter is below.

-Hilary

September 3, 2020

Liron Pantanowitz M.D.
 Director, Division of Anatomic Pathology
 A. James French Professor of Pathology
 Department of Pathology & Clinical Labs
 NCRC Bldg 35, 2800 Plymouth Road
 University of Michigan, Ann Arbor, MI
 Tel 734-232-7318
lironp@med.umich.edu

Dear Dr. Pantanowitz,

I grant you permission to use the article by Pantanowitz L, Michelow P, Hazelhurst S, et al, entitled "A Digital Pathology Solution to Resolve the Tissue Floater Conundrum" (*Arch Pathol Lab Med.* 2020. doi:10.5858/arpa.2020-0034-OA published online ahead of print July 15, 2020). The reprinted article will be used in full as part of your doctoral dissertation, submitted to the University of Witwatersrand, Johannesburg, South Africa.

Full credit to the *Archives of Pathology & Laboratory Medicine* must be given in the reprinted material. Please add "Reprinted with permission. Copyright 2020. College of American Pathologists." This can be added just below or above the full citation.

If listing the article anywhere, you should include the following permission note: "Reprinted from [Author(s), Article title. *Arch Pathol Lab Med.* Year;vol(issue no.): inclusive pages] with permission from *Archives of Pathology & Laboratory Medicine*. Copyright 2020. College of American Pathologists."

This permission is exclusive to this request. Additional usage of any printed or electronic material for which the *Archives of Pathology & Laboratory Medicine* owns the copyright would require permission from the editorial office.

Sincerely,

Donna Hansel, MD, PhD
 Interim Editor in Chief
Archives of Pathology & Laboratory Medicine

Ethics Clearance Certificate

University of Witwatersrand: Human Research Ethics Committee [Medical] clearance
M191003

UNIVERSITY OF THE
WITWATERSRAND
JOHANNESBURG



R14/49 Dr Liron Pantanowitz

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M191003 MED 19-07-044

NAME: Dr Liron Pantanowitz
(Principal Investigator)

DEPARTMENT: Anatomical Pathology
University of Pittsburgh Medical Center (UPMC), USA

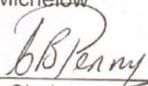
PROJECT TITLE: Digital pathology and artificial Intelligence: Feasibility
for Clinical Practice

DATE CONSIDERED: 25/10/2019

DECISION: Approved unconditionally

CONDITIONS:

SUPERVISOR: Dr Pamela Michelow

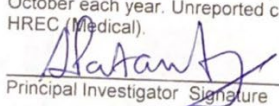
APPROVED BY: 
Dr C Penny, Chairperson, HREC (Medical)

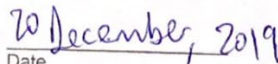
DATE OF APPROVAL: 11/12/2019

This clearance certificate is valid for 5 years from date of approval. Extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office Secretary in Room 301, Third floor, Faculty of Health Sciences, Phillip Tobias Building, 29 Princess of Wales Terrace, Parktown, 2193, University of the Witwatersrand. I/we fully understand the conditions under which I am/we are authorized to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated, from the research protocol as approved, I/we undertake to resubmit the application to the Committee. **I agree to submit a yearly progress report.** The date for annual re-certification will be one year after the date of convened meeting where the study was initially reviewed. In this case, the study was initially reviewed October and will therefore be due in the month of October each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).


Principal Investigator Signature


Date

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES



Liron Pantanowitz <lpantanowitz@gmail.com>

M191003

3 messages

Iain Burns <iain.burns@wits.ac.za>

Wed, Nov 11, 2020 at 5:38 AM

To: "lpantanowitz@gmail.com" <lpantanowitz@gmail.com>



Dr Pantanowitz

Thanks for your letter to the Ethics Committee dated 12 October and received on 30 October.

Congratulations on an impressive publications output.

Your current clearance is valid until 10/12/2024. Annual renewal is normally only requested at the behest of US project funders, so if you are not in that position, your annual reports to the Ethics Committee are normally just for noting.

Iain Burns

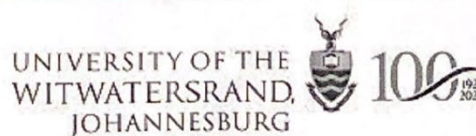
Iain Burns

| Research Office

E : iain.burns@wits.ac.za

T : +27 11 717 1252

W : www.wits.ac.za



3rd Floor, Room 306, Philip Tobias Building, Cnr York Road and 29
Princess of Wales Terrace,
Faculty of Health Sciences, Parktown, Johannesburg

