# iQC: machine-learning-driven prediction of surgical procedure uncovers systematic confounds of cancer whole slide images in specific medical centers

Andrew J. Schaumberg[1,a,*], Michael S. Lewis[2,3,4,b], Ramin Nazarian[5,6,7], Ananta Wadhwa[5], Nathanael Kane[5,8], Graham Turner[1], Purushotham Karnam[1], Poornima Devineni[1], Nicholas Wolfe[1], Randall Kintner[1], Matthew B. Rettig[5,7,9,10,c], Beatrice S. Knudsen[11,12,d], Isla P. Garraway[5,7,13,e], and Saiju Pyarajan[1,f]

[1]Center for Data and Computational Sciences, VA Boston Healthcare System, Boston, MA, USA
[2]Department of Pathology, VA Greater Los Angeles Medical Center, CA, USA
[3]Departments of Pathology and Medicine, The Cedars-Sinai Medical Center, Los Angeles, CA, USA
[4]Center for Cancer Research and Therapeutic Development, Clark Atlanta University, GA, USA
[5]Veterans Affairs (VA) Greater Los Angeles (GLA) Healthcare System, Los Angeles, CA, USA

[6]Department of Medical Research, VA GLA Healthcare System, Los Angeles, CA, USA
[7]Department of Urology, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA
[8]Department of Radiation Oncology, David Geffen School of Medicine at the University of California Los Angeles (UCLA), Los Angeles, CA, USA
[9]Department of Medicine, Division of Hematology-Oncology, VA GLA, Los Angeles, CA, USA
[10]Department of Medicine, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA
[11]Huntsman Cancer Institute BMP core, University of Utah, Salt Lake City, Utah 84108, USA
[12]Department of Pathology, University of Utah, Salt Lake City, Utah 84108, USA
[13]UCLA Jonsson Comprehensive Cancer Center, Los Angeles, CA, USA
[a]AJS https://orcid.org/0000-0001-7556-9208

[b]MSL https://orcid.org/0000-0002-1892-4127
[c]MBR https://orcid.org/0000-0002-7394-3056
[c]BSK https://orcid.org/0000-0002-7589-7591
[e]IPG https://orcid.org/0000-0002-1129-4636

[f]SP https://orcid.org/0000-0002-9047-3762

[*]Correspondence: andrew.schaumberg@va.gov or ajs625@cornell.edu

September 19, 2023

**Abstract**

**Problem:** The past decades have yielded an explosion of research using artificial intelligence for cancer detection and diagnosis in the field of computational pathology. Yet, an often unspoken assumption of this research is that a glass microscopy slide faithfully represents the underlying disease. Here we show systematic failure modes may dominate the slides digitized from a given medical center, such that neither the whole slide images nor the glass slides are suitable for rendering a diagnosis.

**Methods:** We quantitatively define high quality data as a set of whole slide images where the type of surgery the patient received may be accurately predicted by an automated system such as ours, called "iQC". We find iQC accurately distinguished biopsies from nonbiopsies, e.g. prostatectomies or transurethral resections (TURPs, a.k.a. prostate chips), only when the data qualitatively appeared to be high quality, e.g. vibrant histopathology stains and minimal artifacts. Crucially, prostate needle biopsies appear as thin strands of tissue, whereas prostatectomies and TURPs appear as larger rectangular blocks of tissue. Therefore, when the data are of high quality, iQC (i) accurately classifies pixels as tissue, (ii) accurately generates statistics that describe the distribution of tissue in a slide, and (iii) accurately predicts surgical procedure.

**Results:** While we do not control any medical center's protocols for making or storing slides, we developed the iQC tool to hold all medical centers and datasets to the same objective standard of quality. We validate this standard across five Veterans

Affairs Medical Centers (VAMCs) and the Automated Gleason Grading Challenge (AGGC) 2022 public dataset. For our surgical procedure prediction task, we report an Area Under Receiver Operating Characteristic (AUROC) of 0.9966-1.000 at the VAMCs that consistently produce high quality data and AUROC of 0.9824 for the AGGC dataset. In contrast, we report an AUROC of 0.7115 at the VAMC that consistently produced poor quality data. An attending pathologist determined poor data quality was likely driven by faded histopathology stains and protocol differences among VAMCs. Specifically, this VAMC produced slides entirely by hand, whereas all other VAMCs leveraged automated methods to produce slides.

**Conclusion:** Our surgical procedure prediction AUROC may be a quantitative indicator positively associated with high data quality at a medical center or for a specific dataset. To produce high quality data, we recommend producing slides using robotics or other forms of automation whenever possible. We recommend scanning slides digitally before the glass slide has time to develop signs of age, e.g faded stains and acrylamide bubbles. To our knowledge, iQC is the first automated system in computational pathology that validates data quality against objective evidence, e.g. surgical procedure data available in the EHR or LIMS, which requires zero efforts or annotations from anatomic pathologists.

# 1 Introduction

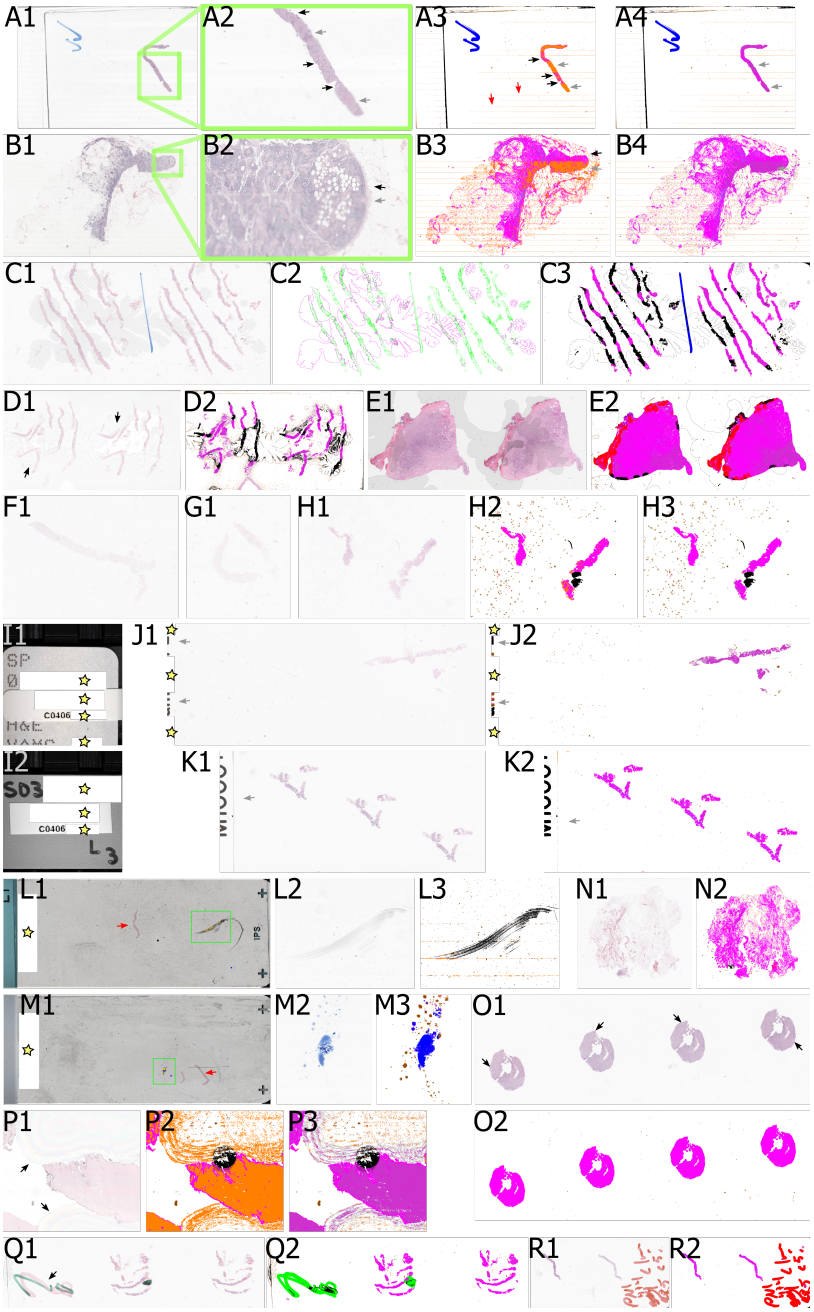
Over the past several years in the field of computational pathology[1], automated methods to assess data quality have tended to focus on excluding confounded regions or measuring the negative impact of poor quality data. The field has benefited from specialized tools such as blur detection[2], pen detection[3], and fragment detection[4] – as well as general frameworks for quality control[5, 6]. The negative effect of whole slide image artifacts on downstream computational pathology methods has been benchmarked[7, 8].

If an automated system is not used to segment out artifacts or other confounds that would lower data quality, computational pathology pipelines may implicitly or explicitly have methods to exclude artifacts. For instance, our earlier work to predict SPOP mutation in prostate cancer was engineered to focus on diagnostically salient regions enriched in predicted subtypes of nuclei, which implicitly avoids pen and background[9]. Later, in what would form the basis of Paige Prostate, Campanella and colleagues used weakly supervised learning and a recurrent neural network to identify suspect foci of cancer in whole slide images, which explicitly used Otsu's method[10] to exclude background and through machine learning at scale may implicitly learn to avoid some artifacts[11]. Shortly thereafter, Lu and colleagues used weakly supervised learning with an attention mechanism for renal cancer subtyping, which explicitly used thresholding to exclude background, while their attention mechanism was shown to exclude normal morphology and some artifacts[12].

Unfortunately, in a large independent validation, Perincheri and colleagues noted "Areas for improvement were identified in Paige Prostate's handling of poor quality scans", which may suggest weakly supervised learning may benefit from rigorous quality control as a preprocessing step[13]. Quality control may flag poor quality slides for manual review, exclude artifacts, or take other actions before Paige Prostate or other downstream processing occurs.

Inspired by our early work that noted different surgical procedures may impact the distribution of tissue in a slide and deep learning performance[14], we developed the hypothesis underlying our iQC tool. Specifically, for high quality data, a quality control system should be able to accurately count the number of tissue pixels in a slide, describe the distribution of tissue in a slide, and therefore predict what kind of surgical procedure was used to excise the tissue present in the slide (Fig 1A1,C1,D1,F1,Q1 are biopsy examples versus Fig 1B1,L2,N1,O1,P1 are nonbiopsy examples). Thus for poor quality data, our hypothesis is that tissue may not be accurately measured and surgical procedure may not be

accurately predicted.




*(Caption on next page.)*

## 2 Results

We found a batch effect, where a subset of data accounted for most slides in iQC's "fail_all_tissue" category (Fig 2A1). iQC defines ten quality control categories (Sec S1.1). Specifically, we found Institution $\beta$ accounted for most of the "fail_all_tissue" slides (Fig 2A2). An anatomic pathologist recommended all slides from Institution $\beta$ fail

**Fig 1.** Representative histopathology images in our study. **A1**: a whole slide image of a prostate needle biopsy at low magnification. **A2**: High magnification of prostate needle biopsy from A1, showing horizontal bands of systematic blur (gray arrows), in contrast to bands of visually sharp pixels where glands are visible (black arrows). **A3**: iQC's quality control mask ($i_{mask_{raw}}$) showing the type of each pixel – background pixels are in white, blue pen pixels (which draw two blue check marks rotated $90^\circ$) are in blue, the edge of the slide is in black, tissue pixels are in magenta (black arrows), "suspect" tissue pixels are orange (gray arrows), and "suspect" pixels that form horizontal bands (red arrows) may suggest the quality of this whole slide image suffers from systematic blur. **A4**: iQC's quality control mask ($i_{mask_{inferred}}$) shows machine learning infers "suspect" tissue pixels as tissue (dark magenta at graw arrows), so all tissue in the slide may be accurately measured for biopsy/nonbiopsy prediction. **B1**: A pelvic lymph node at low magnification where systematic blur may be difficult to perceive. **B2**: Higher magnification plainly shows a horizontal band of systematic blur (gray arrow) compared to visually sharp pixels (black arrow). **B3**: This $i_{mask_{raw}}$ shows sharp pixels are assigned the "tissue" type as indicated in magenta (black arrow), while systematically blurred pixels have the "suspect" type as indicated in orange (gray arrow). **B4**: Machine learning infers "suspect" tissue pixels as tissue, which is shown as a dark magenta (gray arrow). **C1**: A prostate needle biospy, where the slide shows signs of age. **C2**: iQC's quality control mask $i_{mask_{edge}}$ outlines in magenta these signs of age, i.e. large bubbles from degraded acrylamide incompletely holding the coverslip to the glass slide. Tissue and pen are outlined in green. **C3**: $i_{mask_{inferred}}$ shows some tissue pixels in magenta, while other tissue pixels are shown in black, which may loosely correspond to which tissue is most confounded by bubbles. Bubble edges are shown in black. **D1**: This slide shows signs of age through refractive dispersion that causes a rainbow effect (black arrows), in addition to bubbles. **D2**: Like C3, $i_{mask_{inferred}}$ shows tissue in magenta and the most age-counfounded pixels in black. **E1**: The Cancer Genome Atlas (TCGA) is a public dataset. There are bubbles throughout slide TCGA-QU-A6IM-01Z, which may underline the value of automated quality control for public datasets. **E2**: $i_{mask_{inferred}}$ shows bubble edges or bubble-confounded regions in black, regions with blood/erythrocytes in red, and regions with blue marker in blue. **F1,G1,H1**: all these slides have faded histopathology stains. A pathologist deemed these slides unsuitable for diagnosis, in accordance with iQC's stain strength statistics that are weak for these slides. **H2,H3**: $i_{mask_{raw}}$ and $i_{mask_{inferred}}$, respectively, which show points of debris as brown spots, threads of debris in black, and thread-confounded tissue in black. **I1**: A whole slide image thumbnail showing identifiers such as surgical pathology number "SP..." that may be printed on the glass slide, along with other identifiers. A coded external ID "C..." may be applied as a sticker on top to redact some or all of these identifiers. We indicate our redactions to this image with stars. **I2**: A whole slide image thumbnail that shows an accession number "S03..." and a coded external ID "C...". We remove all thumbnails from slides because no identifiers are allowed in research data. Our redactions are indicted with stars. **J1,J2,K1,K2**: Depending on how the glass slide is physically aligned during scanning, text or potentially identifiers on the slide (see I1,I2) may be scanned in the whole slide image at high resolution (at gray arrows, stars for redactions). iQC flags for manual review slides having such markings because patient names or other identifiers are not allowed in research data. **L1**: The thumbnail indicates a black scuff artifact was scanned at high resolution (green box) – missing the prostate needle biopsy (red arrow). **L2,L3**: There is no human tissue scanned at high resolution in this slide, only the black artifact. **M1**: The thumbnail indicates a blue pen mark was scanned at high resolution (green box) – missing the prostate needle biopsy (red arrow). **M2,M3**: There is no human tissue scanned at high resolution in this slide, only the blue artifact. **N1,N2**: Iliac bone in our dataset. **O1,O2**: Colon polypectomy in our dataset, with colonic crypts visible (black arrows). **P1**: Slide with faded stain and more extensive refractive dispersion (black arrows) than D1. **P2**: Due to faded stain and slide age, many pixels are have the "suspect" type in $i_{mask_{raw}}$ (orange). **P3**: Underlining the importance of iQC's machine learning to infer pixel types, these pixels are re-typed as tissue (dark magenta) in $i_{mask_{inferred}}$. Other suspect pixels are inferred as background (gray). **Q1,Q2**: Green pen over red pen (black arrow) typed as green or black in $i_{mask_{inferred}}$. **R1,R2**: iQC detects red pen.

quality control and be remade (Fig 2B2). In the public dataset for the Automated Gleason Grading Challenge 2022[15], we found zero slides failed iQC's quality control, using the same quality control code, configuration, and thresholds (Fig 2A3).

After manually reviewing dozens of cases in detail from Institution $\alpha$, we determined the equations and parameters for iQC's surgical procedure (i.e. biopsy/nonbiopsy) predictor (Sec S1.2). In this way, iQC achieved AUROC of 0.9966 for the biopsy/nonbiopsy prediction task (Fig 3A). Testing this on all other VAMC data, we found AUROC substantially dropped to AUROC of 0.8346 (Fig 3B). Testing only on Institution $\beta$ data, we found even lower AUROC of 0.7115 (Fig 3C). Testing on the Institutions that were neither $\alpha$ nor $\beta$, we found AUROC of 1.000 (Fig 3D). We concluded Institution $\beta$ drove the drop in AUROC on non-Institution-$\alpha$ data (Fig 3B).

To test how well iQC generalized to unseen data, and more specifically to non-Veteran data, we evaluated iQC's biopsy/nonbiopsy predictor on AGGC2022 data[15]. We found an AUROC of 0.9824 (Fig 3E). This may suggest iQC generalizes well to unseen data, external datasets, and non-Veteran data. An example of a nonbiopsy that iQC mistakes for a biopsy is shown in Figure 3F1, where the ectomy tissue is cut into a long a thin strip. Biopsies tend to be long and thin.

# 3 Methods

This study was approved by the Institutional Review Board at the VA Boston Healthcare System.

iQC generates interpretable statistics to calculate a score for biopsy/nonbiopsy prediction. iQC has a multistep pipeline, including (1) Otsu thresholds, (2) initial pixel typing, (3) debris detection, (4) edge detection, (5) blur artifact detection, (6) black mark detection, (7) pen detection, (8) stain strength calculations and pen extension, (9) write $i_{mask_{raw}}$ mask, (10) blur artifact orientation detection, (11) writing of $i_{mask_{mean}}$ that mixes base image with $i_{mask_{raw}}$, (12) machine-learning-driven inference of "suspect" type pixels to other types e.g. pen, tissue, and background, (13) write $i_{mask_{inferred}}$ mask, (14) biopsy/nonbiopsy prediction, (15) ridge detection and age-related bubble detection, (16) write $i_{mask_{edge}}$ mask, (17) barcode and text detection for PHI/PII risks, (18) close-out timing statistics.

iQC's biopsy/nonbiopsy predictor is a function that generates a score ($y(i_{m_i})$ in Eqn 1) between 0 and 1000000, with low numbers favoring a biopsy and high numbers favoring a nonbiopsy, e.g. prostatectomy or TURP, where for brevity we denote $i_{mask_{inferred}}$ as $i_{m_i}$:

$$y(i_{m_i}) = min(1000000, G_{narrow}(i_{m_i}) \times G_{area}(i_{m_i}) \times G_{long}(i_{m_i}) + G_{adipose}(i_{m_i})) \quad (1)$$

Each $G(...)$ is a Gompertz function[16], further discussed in the supplement (Sec S1.2).

Computational software and hardware (Sec S1.3), as well as whole slide image details (Sec S1.4), are detailed in the supplement.

# 4 Discussion

To our knowledge, we are the first to define quality in terms of objective ground truth data, e.g. data are high quality if the AUROC is close to 1.0 for a surgical procedure prediction task, or other objective ground truth available from a Laboratory Information Management System (LIMS). For iQC, AUROC is close to 1.0 for all datasets and VAMCs exception Institution $\beta$. Institution $\beta$ has AUROC of 0.71 and poor quality data.

Institution $\beta$ slides were made in the years 2000-2007. The slides were over a decade old when they were scanned in 2023. The slides from the other VAMCs were made in the years 2003-2021. We believe this suggests there are myriad technical factors beyond calendar age that contribute to how over time slides show signs of age, e.g. faded stain and acrylamide bubbles. Technical factors may include the choice of hematoxylin and eosin (H&E) stains as well as storage conditions of the slides.

iQC detected the faded stain and assigned many Institution $\beta$ slides to the "fail_all_slide" quality control category to indicate these slides are not suitable for diagnosis. An anatomic pathologist later reviewed the slides and recommended all Institution $\beta$ slides fail quality control and be remade.

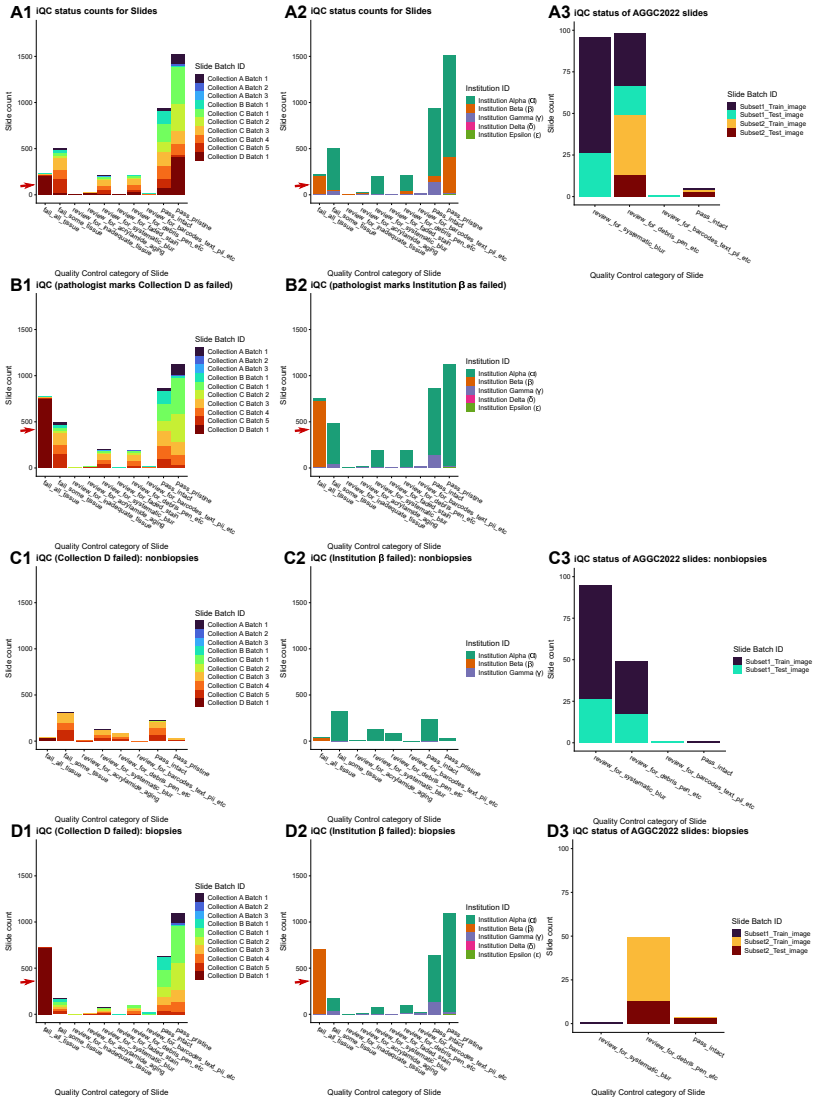



**Fig 2.** iQC quality control category counts for slides in our study. **A1**: Initial data collection. A "collection" is a set of data, e.g. all slides from a conference room or a prior study. A "batch" corresponds to a subset of slides, e.g. the first box of slides from the conference room. The next box would be batch 2. We noticed "Collection D Batch 1" accounted for most slides in iQC's "fail_all_tissue" category (red arrow), which indicates these slides have faded stain, so the entire slide is unsuitable for diagnosis. **A2**: Coloring A1 according to the VAMC/Institution that made the slide, it appears Institution Beta ($\beta$) accounts for most "fail_all_tissue" slides, which is a batch effect. **A3**: Zero AGGC2022 slides fail iQC's quality control, suggesting this dataset is high quality[15]. Most slides are "review...", which is a warning. Many slides have tissue debris. **B1,B2**: A pathologist reviewed slides from Institution $\beta$. He concluded they were poor quality and unsuitable for diagnosis. He recommended all slides from Institution $\beta$ fail quality control (red arrow). **C1,C2,C3**: Quality control categories for slides iQC predicted to be nonbiopsies, e.g. prostatectomies or TURPs. **D1,D2,D3**: Quality control categories for slides iQC predicted to be biopsies of any sort, e.g. prostate needle biopsies or liver biopsies.

iQC's definition of quality in terms of objective ground truth data differs from some prior approaches to quality control in digital pathology. In 2019, Schaumberg and Fuchs defined quality control in terms of numeric parameters or rules that were hand-engineered[6]. To

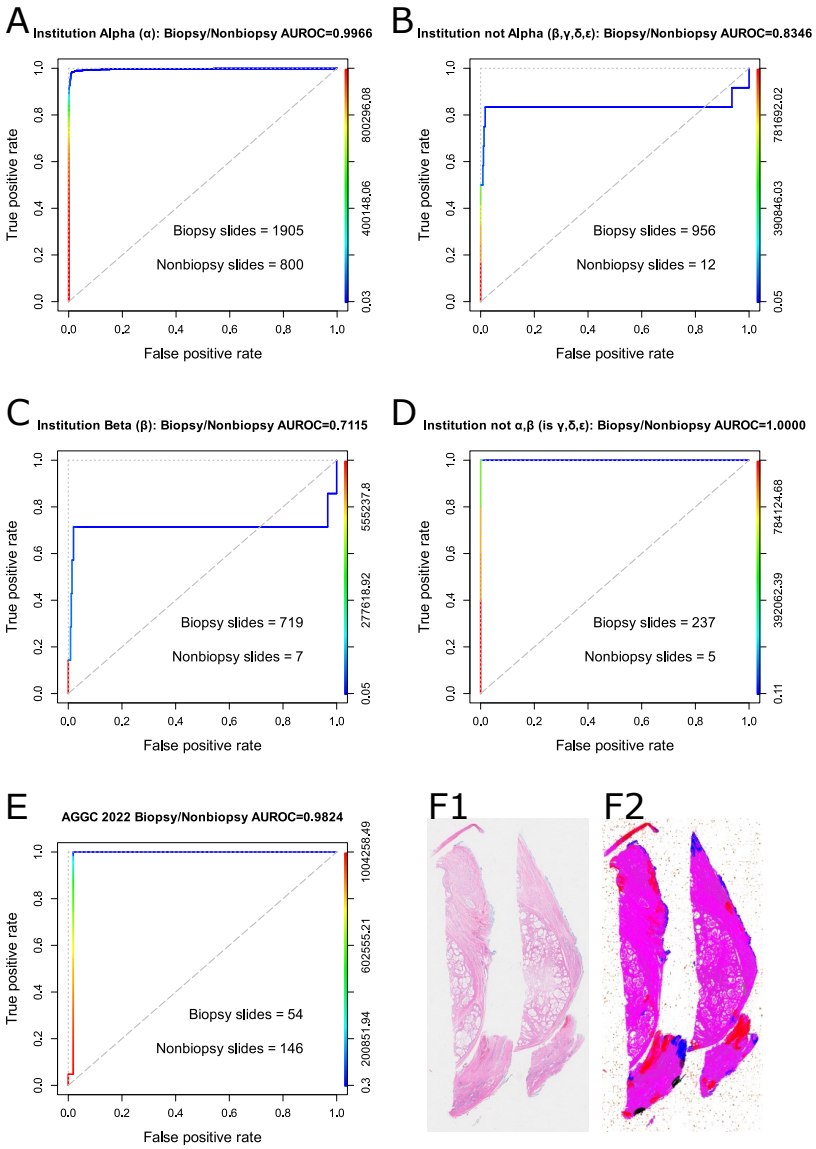



**Fig 3.** iQC biopsy/nonbiopsy prediction AUROC for various datasets and subsets in our study. **A**: AUROC for Institution Alpha ($\alpha$), which made most VAMC slides. We trained/tuned biopsy/nonbiopsy predictor on a subset of this dataset, so AUROC is high. **B**: AUROC for all institutions that are not $\alpha$, i.e. $\beta, \gamma, \delta, \epsilon$. AUROC is much lower. There are few nonbiopsy slides. **C**: AUROC for $\beta$, which is strikingly low. We believe this is because Institution $\beta$ provided poor quality slides that were old and had faded stain. iQC is not able to accurately identify which pixels are tissue and cannot distinguish biopsies from nonbiopsies. **D**: AUROC for other VAMCs (i.e. $\gamma, \delta, \epsilon$) is high but this is an underpowered test because there are only 5 nonbiopsies. **E**: AUROC on the public AGGC2022 dataset[15] is high, indicating iQC's biopsy/nonbiopsy predictor may generalize well to unseen data. **F1,F2**: iQC mistakenly classified this prostatectomy sample as a needle biopsy, perhaps because the distribution of tissue is long and thin, very loosely like a needle biopsy.

refine the subjective nature of quality control, HistoQC validated quality control in terms of pathologist concordance[17]. HistoROI claimed to improve upon HistoQC through human-in-the-loop training and a deep learning system on annotated tiles, although this quality control depends on manual annotations from pathologists and their subjective

interpretation of the morphology[18].

We believe iQC is a novel approach to quality control because it is rooted in objective ground truth. Highly curated surgical procedure or other coded data from the LIMS or Electronic Health Record (EHR) are readily available. This approach may scale well because surgical procedure annotations are at the whole slide level, rather than at the region of interest (ROI) or pixel level. Slide-level annotations drive the scalability of Campanella[11], Lu[12], and other weakly supervised learning pipelines. Still, iQC provides quality control masks for per-pixel semantic segmentation, e.g. $i_{mask_{inferred}}$, to assist pathologists and downstream AI pipelines in distinguishing artifacts such as pen or blur from tissue in the whole slide image.

Proceeding from some of our earlier work that showed how tractable search is in computational pathology[19, 20], we framed quality control as a search problem at the pixel level. Rather than train artifact-specific classifiers to detect blur[2], pen[3], or coverslip breaks[5], iQC uses machine learning to compute how similar a "suspect" pixel is to other "nonsuspect" pixel types (e.g. pen, tissue, background, etc see Section S1.5). We believe this allows iQC to define relatively simple rules for the appearance of different pixel types, and extend these rules using machine-learning-driven inferences (Fig 1P1-P3), to achieve high AUROC performance across VA and AGGC datasets (Fig 3).

# 5 Conclusion

Our iQC pipeline found a batch effect from a medical center that provided old slides to scan and made the slides using a manual process that differed from the repeatable automated processes used at other medical centers. Moreover, iQC provides type information for each pixel, uses pixel type information to predict surgical procedure, and provides an overall AUROC for surgical procedure prediction performance that positively corresponds to the overall quality of data produced at a medical center.

We find high AUROC for all datasets and medical centers except the one medical center with old slides and manual procedures. At this medical center ($\beta$) AUROC is correspondingly much lower and histopathology stains are faded. Because iQC separates biopsies from nonbiopsies in mixed incoming datasets, we believe iQC may be especially valuable for downstream studies where only biopsies may be included a study, to the exclusion of all other surgical procedures, i.e. prostatectomies, TURPs, colonic polypectomies, etc.

To our knowledge, we present the first quality control pipeline for histopathology validated to objective ground truth data, specifically surgical procedure. Following this approach, a hospital may apply our quality control pipeline and validate against surgical procedure data in their LIMS or EHR, without requiring effort or annotations from pathologists. We encourage broad adoption of such scalable quality control pipelines in digital pathology and computational pathology pipelines.

# 6 Acknowledgements

This work was funded through a Prostate Cancer Foundation grant to MBR, BSK, IPG, and SP. Authors thank Mark Hewitt and Nicholas Burns for high performance computing support. AJS thanks Mariam Aly for early manuscript discussion. We are grateful to the patients who made this study possible.

# 7 Contributions

Conceptualization: AJS, SP.

Data acquisition: AJS and RK (AGGC slides), AW and NK (VA slides), NW (LIMS biopsy/nonbiopsy).
Data curation: AJS, MSL.
Data transfer and management: AJS, RN, AW, NK, GT, PK, PD, NW, RK.
Methodology, software, validation, formal analysis, investigation, visualization, writing (original draft): AJS.
Funding acquisition: MBR, BSK, IPG, SP.
Project administration: AJS, MSL, RN, AW, NK, MBR, BSK, IGP, SP.
Resources (pathology) and discussion: MSL, RN, AW, NK, BSS, IPG.
Resources (computational) and discussion: GT, PK, PD, RK, SP.
Supervision (pathology): MSL, RN, IPG.
Supervision (computational): MSL, SP.
Writing (editing): AJS.
Writing (reviewing): AJS, IPG, PK.

# 8 Ethics Declaration and Conflicts of Interest

The author(s) declare they have no competing interests.

# References


[1] Thomas Fuchs and Joachim Buhmann. "Computational pathology: challenges and promises for tissue analysis." In: *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society* 35.7-8 (Oct. 2011), pp. 515–530. ISSN: 1879-0771. DOI: 10.1016/j.compmedimag.2011.02.006. URL: http://dx.doi.org/10.1016/j.compmedimag.2011.02.006.

[2] Gabriele Campanella et al. "Towards machine learned quality control: A benchmark for sharpness quantification in digital pathology". In: *Computerized Medical Imaging and Graphics*. Advances in Biomedical Image Processing 65 (Apr. 2018), pp. 142–151. ISSN: 0895-6111. DOI: 10.1016/j.compmedimag.2017.09.001. URL: https://www.sciencedirect.com/science/article/pii/S0895611117300800 (visited on 09/12/2023).

[3] Peter J. Schüffler et al. "Overcoming an Annotation Hurdle: Digitizing Pen Annotations from Whole Slide Images". In: *Journal of Pathology Informatics* 12.1 (Jan. 2021), p. 9. ISSN: 2153-3539. DOI: 10.4103/jpi.jpi_85_20. URL: https://www.sciencedirect.com/science/article/pii/S2153353922001316 (visited on 09/12/2023).

[4] Tomé Albuquerque et al. "Quality Control in Digital Pathology: Automatic Fragment Detection and Counting". In: *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. ISSN: 2694-0604. July 2022, pp. 588–593. DOI: 10.1109/EMBC48229.2022.9871208.

[5] Andrew Janowczyk et al. "HistoQC: An Open-Source Quality Control Tool for Digital Pathology Slides". In: *JCO Clinical Cancer Informatics* 3 (Dec. 2019). Publisher: Wolters Kluwer, pp. 1–7. DOI: 10.1200/CCI.18.00157. URL: https://ascopubs.org/doi/full/10.1200/CCI.18.00157 (visited on 09/12/2023).

[6] Andrew Schaumberg and Thomas Fuchs. "Identifying regions of interest from whole slide images". Provisional application filed 2019-08-24, utility patent awarded 2021-03-30. 10963673 (New York, NY). Mar. 2021. URL: https://patentscope.wipo.int/search/en/detail.jsf?docId=WO2021041338.

[7] Alexander I. Wright et al. "The Effect of Quality Control on Accuracy of Digital Pathology Image Analysis". In: *IEEE Journal of Biomedical and Health Informatics* 25.2 (Feb. 2021). Conference Name: IEEE Journal of Biomedical and Health Informatics, pp. 307–314. ISSN: 2168-2208. DOI: 10.1109/JBHI.2020.3046094.

[8] Birgid Schömig-Markiefka et al. "Quality control stress test for deep learning-based diagnostic model in digital pathology". In: *Modern Pathology* 34.12 (Dec. 2021), pp. 2098–2108. ISSN: 0893-3952. DOI: 10.1038/s41379-021-00859-x. URL: https://www.sciencedirect.com/science/article/pii/S0893395222003702 (visited on 09/12/2023).

[9] Andrew Schaumberg, Mark Rubin, and Thomas Fuchs. "H&E-stained Whole Slide Deep Learning Predicts SPOP Mutation State in Prostate Cancer". In: *bioRxiv* (July 2016), p. 064279. DOI: 10.1101/064279. URL: http://dx.doi.org/10.1101/064279.

[10] Nobuyuki Otsu. "A Threshold Selection Method from Gray-Level Histograms". In: *IEEE Transactions on Systems, Man, and Cybernetics* 9.1 (Jan. 1979). Conference Name: IEEE Transactions on Systems, Man, and Cybernetics, pp. 62–66. ISSN: 2168-2909. DOI: 10.1109/TSMC.1979.4310076.

[11] Gabriele Campanella et al. "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images". eng. In: *Nature Medicine* 25.8 (2019), pp. 1301–1309. ISSN: 1546-170X. DOI: 10.1038/s41591-019-0508-1.

[12] Ming Y. Lu et al. "Data-efficient and weakly supervised computational pathology on whole-slide images". en. In: *Nature Biomedical Engineering* (Mar. 2021), pp. 1–16. ISSN: 2157-846X. DOI: 10.1038/s41551-020-00682-w. URL: https://www.nature.com/articles/s41551-020-00682-w (visited on 03/02/2021).

[13] Sudhir Perincheri et al. "An independent assessment of an artificial intelligence system for prostate cancer detection shows strong diagnostic accuracy". en. In: *Modern Pathology* 34.8 (Aug. 2021). Number: 8 Publisher: Nature Publishing Group, pp. 1588–1595. ISSN: 1530-0285. DOI: 10.1038/s41379-021-00794-x. URL: https://www.nature.com/articles/s41379-021-00794-x (visited on 09/14/2023).

[14] Andrew Schaumberg et al. "DeepScope: Nonintrusive Whole Slide Saliency Annotation and Prediction from Pathologists at the Microscope". In: Sept. 2016. ISBN: 978-3-319-67834-4. DOI: https://doi.org/10.1007/978-3-319-67834-4_4. URL: https://www.ncbi.nlm.nih.gov/pubmed/29601065.

[15] Xinmi Huo et al. *Comprehensive AI Model Development for Gleason Grading: From Scanning, Cloud-Based Annotation to Pathologist-AI Interaction*. en. SSRN Scholarly Paper. Rochester, NY, July 2022. DOI: 10.2139/ssrn.4172090. URL: https://papers.ssrn.com/abstract=4172090 (visited on 09/16/2023).

[16] Benjamin Gompertz. "On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies". In: *Philosophical Transactions of the Royal Society of London* 115 (Dec. 1825). Publisher: Royal Society, pp. 513–583. DOI: 10.1098/rstl.1825.0026. URL: https://royalsocietypublishing.org/doi/10.1098/rstl.1825.0026 (visited on 09/17/2023).

[17] Andrew Janowczyk, Patrick Leo, and Mark A Rubin. "Clinical deployment of AI for prostate cancer diagnosis". en. In: *The Lancet Digital Health* 2.8 (Aug. 2020), e383–e384. ISSN: 2589-7500. DOI: 10.1016/S2589-7500(20)30163-1. URL: http://www.sciencedirect.com/science/article/pii/S2589750020301631 (visited on 08/03/2020).

[18] Abhijeet Patil et al. "Efficient quality control of whole slide pathology images with human-in-the-loop training". In: *Journal of Pathology Informatics* 14 (Jan. 2023), p. 100306. ISSN: 2153-3539. DOI: 10.1016/j.jpi.2023.100306. URL: https://www.sciencedirect.com/science/article/pii/S2153353923001207 (visited on 09/15/2023).

[19] Andrew J. Schaumberg et al. "Interpretable multimodal deep learning for real-time pan-tissue pan-disease pathology search on social media". en. In: *Modern Pathology* 33.11 (Nov. 2020), pp. 2169–2185. ISSN: 1530-0285. DOI: 10.1038/s41379-020-0540-1. URL: https://www.nature.com/articles/s41379-020-0540-1/ (visited on 08/11/2021).

[20] Chengkuan Chen et al. "Fast and scalable search of whole-slide images via self-supervised deep learning". en. In: *Nature Biomedical Engineering* 6.12 (Dec. 2022). Number: 12 Publisher: Nature Publishing Group, pp. 1420–1434. ISSN: 2157-846X. DOI: 10.1038/s41551-022-00929-8. URL: https://www.nature.com/articles/s41551-022-00929-8 (visited on 09/18/2023).

[21] Charles R. Harris et al. "Array programming with NumPy". en. In: *Nature* 585.7825 (Sept. 2020). Number: 7825 Publisher: Nature Publishing Group, pp. 357–362. ISSN: 1476-4687. DOI: 10.1038/s41586-020-2649-2. URL: https://www.nature.com/articles/s41586-020-2649-2 (visited on 09/18/2023).

[22] Tobias Sing et al. "ROCR: visualizing classifier performance in R." In: *Bioinformatics (Oxford, England)* 21.20 (Oct. 2005), pp. 3940–3941. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bti623. URL: http://dx.doi.org/10.1093/bioinformatics/bti623.

[23] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. URL: https://ggplot2.tidyverse.org (visited on 09/18/2023).

[24] Adam Goode and Mahadev Satyanarayanan. *A Vendor-Neutral Library and Viewer for Whole-Slide Images*. June 2008. URL: http://reports-archive.adm.cs.cmu.edu/anon/2008/abstracts/08-136.html.

[25] Adam Goode et al. "OpenSlide: A Vendor-Neutral Software Foundation for Digital Pathology". In: *Journal of Pathology Informatics* 4.1 (2013), p. 27. DOI: https://doi.org/10.4103/2153-3539.119005. URL: https://www.sciencedirect.com/science/article/pii/S2153353922006484.

[26] Evelyn Fix and Joseph L. Hodges. "Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties". en. In: *USAF School of Aviation Medicine* ADA800276.5028-110633 (1951), pp. 1–24. URL: https://apps.dtic.mil/dtic/tr/fulltext/u2/a800276.pdf.

[27] T Cover and P Hart. "Nearest neighbor pattern classification". In: *IEEE Transactions on Information Theory* 13.1 (Jan. 1967), pp. 21–27. ISSN: 0018-9448. DOI: 10.1109/tit.1967.1053964. URL: http://dx.doi.org/10.1109/tit.1967.1053964.

[28] Irwin Sobel and Gary Feldman. *A 3x3 Isotropic Gradient Operator for Image Processing*. en. Invited talk. 1968.

[29] Irwin Sobel. "An Isotropic 3x3 Image Gradient Operator". In: *Presentation at Stanford A.I. Project 1968* (Feb. 2014). URL: https://www.researchgate.net/publication/239398674_An_Isotropic_3_3_Image_Gradient_Operator.

# Supporting Information

# S1 Supplementary materials and methods

## S1.1 iQC quality control categories

iQC defines ten quality control categories (Fig 2). This provides granualr information for the quality of a slide. These categories are grouped into "fail...", "review...", and "pass..." supercategories. We define each of the ten categories below.

1. fail_all_tissue: iQC suggests all tissue in the slide is not suitable for any diagnostic purpose. The tissue is not suitable for a pathologist to render a diagnosis. The tissue is also not suitable for downstream computational analysis / machine learning / artificial intelligence (AI). Typically, this occurs because the tissue staining is badly faded, e.g. the hematoxylin stain is not visible and only eosin remains. This fading worsens as the slide ages, depending on the storage conditions, the quality of stains used, and perhaps other factors.

2. fail_some_tissue: iQC suggests some of the tissue is not suitable for any diagnostic purpose. This typically occurs if the acrylamide layer that adheres the glass coverslip to the glass slide has aged. Such acrylamide age leads to "window pane breaking" artifacts and bubbles. If these artifacts or bubbles occur over tissue in the slide, generally speaking such occluded tissue is not suitable for a pathologist or AI. These regions should be excluded, while the rest of the unaffected tissue may be retained and used by a pathologist or AI. It may be especially problematic if such artifacts or bubbles occlude the only malignant foci in the slide, or other foci of disease. Such occlusion may change the diagnosis, depending on whether or not disease foci are include or excluded. For this reason, great care should be taken when using fail_some_tissue slides.

3. review_for_inadequate_tissue: iQC suggests there is very little tissue in the slide. This slide should be manually reviewed by an expert to determine if sufficient tissue exists for either a pathologist to diagnose a disease or an AI to analyze.

4. review_for_acrylamide_aging: iQC suggests there may be evidence of acrylamide aging in the slide, e.g. window pane breaking artifacts or bubbles. The evidence is not strong, so manual expert review is recommended.

5. review_for_systematic_blur: iQC suggests there may be evidence of a specific type of blur in the slide, which we call systematic blur. Systematic blur is thought to occur when the acrylamide layer has aged such that the glass coverslip is not securely adhered to the glass slide, so the slide "shakes in place" while the slide is being scanned, and this shaking is such that the scanner's autofocus cannot focus correctly on the slide to get a sharp picture. The result is a band of blurred pixels, e.g. a horizontal band of blurring as the scanner's camera travels left to right to photograph parts of the slide. Systematic blur induces subtle linear artifacts in the background between adjacent passes of the scanner's camera, which may occur when the scanner's software stitches together images. iQC detects these linear artifacts as straight lines of "suspect" pixels. If the number of such linear suspect pixels exceeds a threshold, iQC recommends the slide for manual expert review here. This may be a novel way to detect blur, in that we look at lines in the slide background, rather than directly look for blurry pixels.

6. review_for_faded_stain: iQC suggests there may be evidence of stain fading in the slide, but this evidence is not strong, so manual expert review is recommended.

7. review_for_debris_pen_etc: iQC suggests there may be evidence for debris, pen, or marker in the slide. Manual expert review is recommended. Some care should be taken with these slides, e.g. if pen marks occlude foci of disease, omitting such foci may change the diagnosis.

8. review_for_barcodes_writing_pii_etc: iQC introduces a potentially novel method to detect barcodes or other black structured marks (like text) on a slide. Slides with names printed on them in black text are not de-identified and are not suitable for research purposes. Often, however, the printed text in a slide indicates where the slide was manufactured, rather than indicating PHI/PII of the patient. Manual expert review is recommended.

9. pass_intact: iQC suggests the slide is generally good condition, though there may be some small amount of pen, marker, or debris present in the slide. Automated tools such as iQC may recommend where pen, marker, etc are in the slide so these may be avoided. iQC recommends the slide is otherwise of high quality and is expected to be suitable for both a pathologist and AI.

10. pass_pristine: iQC suggests the entire slide is high quality and can likely be used as-is.

## S1.2 Surgical procedure prediction (biopsy/nonbiopsy)

In Equation 1, $G_{narrow}$ converts to a number between 0 and 100 the approximate measurement of the narrowest region of tissue (Eqn S1), with the intuition that biopsies are narrow and will be close to 0:

$$G_{narrow}(i_{m_i}) = 100e^{-100e^{-0.025*mmts}} \tag{S1}$$

$G_{area}$ converts to a number between 0 and 100 the approximate measurement of the tissue area by summing up the number of tissue type pixels, with the intuition that biopsies tend to involve little tissue and will be close to 0:

$$G_{area}(i_{m_i}) = 100e^{-50e^{-0.000004*(sthp+stnpi)}} \tag{S2}$$

$G_{long}$ converts to a number between 0 and 100 the approximate length-to-width ratio of the tissue, with the intuition that biopsies tend to have a ratio much greater than 1 (so $G_{long}$ will be close to 0) while nonbiopsies tend to have a ratio close to 1 (so $G_{long}$ will be close to 1):

$$G_{long}(i_{m_i}) = 100 - 100e^{-100e^{-2.5*mmctsr}} \tag{S3}$$

$G_{adipose}$ is a correction factor for specimens with an abundance of adipose tissue, to prevent some ectomy/TURP samples that are mostly fat from erroneously being predicted as biopsies (Eqn S4). $G_{adipose}$ converts to a number between 0 and 100000 the approximate measure of how frequently tissue type pixels are adjacent to background type pixels. Background pixels are white/clear/empty in the slide image. Adipose (a.k.a. fat) tissue typically consists of thin strips of stromal tissue to support large globules of fat tissue. The stromal tissue is pink and is counted as tissue pixels, while the fat globules are clear and are counted as background. Therefore in fat, tissue pixels are adjacent to many more background pixels that would occur in a biopsy or in solid blocks of tissue (in most ectomies or TURPs). Fatty ectomies have a high $G_{adipose}$, which is important because fatty ectomies may have very little solid tissue, or only a thin strip of tissue that would otherwise be classified as a biopsy, if it were not for $G_{adipose}$ increasing $y(i_{m_i})$ (Eqn 1).

$$G_{adipose}(i_{m_i}) = 100000e^{-20e^{-10*batr}} \tag{S4}$$

We define $mmts$ (Eqn S1) as the "min median tissue score". Low values of $mmts$ mean the narrowest region is very small, so the tissue may be thin, so $G_{narrow}(i_{m_i})$ approaches 0, to suggest a biopsy. In contrast, high values of $mmts$ mean the narrowest region is much larger, so the tissue is approximately square in shape, $G_{narrow}(i_{m_i})$ approaches 100, to suggest a nonbiopsy.

In Equation S2, we define $sthp$ as the "sum [of] tissue or hematoxylin pixels" and $stnpi$ as "suspect-to-nonsuspect pixels inferred".

We define $mmstsr$ (Eqn S3) as "maxmin max contiguous tissue score ratio". Low values of $mmctsr$ mean the ratio is close to 1 and the tissue is approximately square so $G_{long}(i_{m_i})$ approaches 100, to suggest a nonbiopsy. In contrast, high values of $mmctsr$ mean the ratio may be larger than 1 (e.g. a ratio of 20) and the tissue is approximately ribbon-like so $G_{long}(i_{m_i})$ approaches 0, to suggest a biopsy.

In Equation S4, we define $batr$ as the "background-to-adjacent-tissue ratio".

## S1.3 Computational software and hardware

We implemented iQC in `python 3.9.16`, with imports from `numpy`[21] and `statistics`. Visualization was performed in `R version 4.2.2` and `rstudio 2022.07.2-576`, with plots made via `ROCR`[22] and `ggplot2`[23]. Text processing was performed in `perl 5.32.1`.

For computation, we leveraged the GenISIS supercomputer at the Center for Data and Computational Sciences. iQC can run on a single-CPU system with an amount of CPU RAM approximately triple the file size, e.g. for a 1GB whole slide image, we suggest 3GB CPU RAM. We recommend running iQC on at least 5 CPUs in parallel, ideally 20 CPUs, and 80+ CPUs for best performance. The amount of required CPU RAM scales with the number of parallel CPUs.

## S1.4 Whole slide images

VAMC slides were scanned on a Leica Aperio GT450 scanner, which produces SVS files that we read via `openslide 3.4.1`[24, 25]. AGGC2022 slides were scanned on an Akoya Biosciences scanner, which produces TIFF files that we read via `ImageMagick 6`.

## S1.5 Machine learning

iQC uses machine learning to infer select "suspect" pixel types to other pixel types. Specifically, for a given image, iQC uses the K-nearest neighbors algorithm[26, 27], with k=1 and an L1 norm. A pixel is represented as a four-dimensional value: red channel (0 to 255 integer), green channel, blue channel, and Sobel channel. The Sobel channel is the magnitude of a $3 \times 3$ Sobel operator[28, 29].