Title

Development of a deep learning algorithm for Paneth cell density quantification for inflammatory bowel disease.

Authors

Kang, Liang-I
Sarullo, Kathryn
Marsh, Jon
et al.

Publication Date

2024-11-12

DOI

10.1016/j.ebiom.2024.105440

Peer reviewed

Development of a deep learning algorithm for Paneth cell density quantification for inflammatory bowel disease

Liang-I Kang,[a,d] Kathryn Sarullo,[a,d] Jon N. Marsh,[a,d] Liang Lu,[a] Pooja Khonde,[a] Changqing Ma,[a] Talin Haritunians,[b] Angela Mujukian,[b] Emebet Mengesha,[b] Dermot P. B. McGovern,[b] Thaddeus S. Stappenbeck,[c] S. Joshua Swamidass,[a,*] and Ta-Chiang Liu[a,**]

[a]Department of Pathology & Immunology, Washington University in St. Louis School of Medicine, 660 South Euclid Avenue, Campus Box 8118, Saint Louis, MO, 63110, United States
[b]The F. Widjaja Inflammatory Bowel Disease Institute, Cedars-Sinai Medical Center, 8700 Beverly Blvd., Los Angeles, CA, 90048, United States
[c]Department of Inflammation and Immunity, Cleveland Clinic Foundation, Mail Code NE30, 9500 Euclid Avenue, Cleveland, OH, 44195, United States

eBioMedicine
2024;110: 105440
Published Online xxx
https://doi.org/10.1016/j.ebiom.2024.105440

Summary

Background Alterations in ileal Paneth cell (PC) density have been described in gut inflammatory diseases such as Crohn's disease (CD) and could be used as a biomarker for disease prognosis. However, quantifying PCs is time-intensive, a barrier for clinical workflow. Deep learning (DL) has transformed the development of robust and accurate tools for complex image evaluation. Our aim was to use DL to quantify PCs for use as a quantitative biomarker.

Methods A retrospective cohort of whole slide images (WSI) of ileal tissue samples from patients with/without inflammatory bowel disease (IBD) was used for the study. A pathologist-annotated training set of WSI were used to train a U-net two-stage DL model to quantify PC number, crypt number, and PC density. For validation, a cohort of 48 WSIs were manually quantified by study pathologists and compared to the DL algorithm, using root mean square error (RMSE) and the coefficient of determination (r^2) as metrics. To test the value of PC quantification as a biomarker, resection specimens from patients with CD (n = 142) and without IBD (n = 48) patients were analysed with the DL model. Finally, we compared time to disease recurrence in patients with CD with low versus high DL-quantified PC density using Log-rank test.

Findings Initial one-stage DL model showed moderate accuracy in predicting PC density in cross-validation tests (RMSE = 1.880, r^2 = 0.641), but adding a second stage significantly improved accuracy (RMSE = 0.802, r^2 = 0.748). In the validation of the two-stage model compared to expert pathologists, the algorithm showed good performance up to $RMSE$ = 1.148, r^2 = 0.708. The retrospective cross-sectional cohort had mean ages of 62.1 years in the patients without IBD and 38.6 years for the patients with CD. In the non-IBD cohort, 43.75% of the patients were male, compared to 49.3% of the patients with CD. Analysis by the DL model showed significantly higher PC density in non-IBD controls compared to the patients with CD (4.04 versus 2.99 PC/crypt). Finally, the algorithm quantification of PCs density in patients with CD showed patients with the lowest 25% PC density (Quartile 1) have significantly shorter recurrence-free interval (p = 0.0399).

Interpretation The current model performance demonstrates the feasibility of developing a DL-based tool to measure PC density as a predictive biomarker for future clinical practice.

Funding This study was funded by the National Institutes of Health (NIH).

Keywords: Whole slide images; Crohn's disease; Pathology; Prognosis

*Corresponding author.
**Corresponding author.
E-mail addresses: swamidass@wustl.edu (S.J. Swamidass), ta-chiang.liu@wustl.edu (T.-C. Liu).
[d]These authors contributed equally to this work.

Research in context

Evidence before this study

Paneth cells have been established as a cellular biomarker for clinical outcome in CD by multiple groups. At the time this study was undertaken, there was no other published study that could be found using common search engines such as PubMed or Google Scholar that had developed any computational pathology or deep learning algorithms to quantify Paneth cell density.

Added value of this study

Our study shows a deep learning algorithm that can robustly quantify Paneth cell number and density (Paneth cell per crypt) with enough accuracy to show predictive value in a clinical validation cohort, and can analyse Paneth cells using deep learning algorithms.

Implications of all the available evidence

Paneth cells are an important intestinal epithelial cell type that has been shown to be lost or dysfunctional in many intestinal diseases, and there is increasing interest to use them as a biomarker for disease activity. However, accurate quantification requires substantial time to perform and some specialised training, which currently impedes clinical utility. Our Paneth cell quantification algorithm is a crucial innovation that could aid in translating using Paneth cells as a biomarker in routine surgical pathology clinical practice.

Introduction

Paneth cells (PCs) are specialised intestinal epithelial cells with diverse functions in immunity and host defence.[1–3] In humans, PCs are located at the base of the crypts of Lieberkühn throughout the small intestine and in the proximal portion of the large intestine.[4] PCs are morphologically distinct from other intestinal epithelial cell types and are identified by the presence of numerous cytoplasmic eosinophilic granules, which contain antimicrobial proteins and peptides, such as α-defensins and lysozyme.[5] PCs also secrete niche factors to support the homeostasis of adjacent intestinal stem cells.[6,7] Therefore, PCs are critical regulators of gut innate immunity, and PC defects have been implicated in the pathogenesis of several gut inflammatory disorders, including inflammatory bowel disease (IBD).[8–10]

The pathogenesis of Crohn's disease (CD), a major subtype of IBD, involves host genetic susceptibility and environmental triggers.[11–14] We and others have shown that gene–environmental interactions regulate PC function,[15–20] and more importantly, PC function correlates with clinical outcomes in patients with CD undergoing surgical resection in multiple cohorts.[15–17,21] However, integrating these findings into routine clinical/surgical pathology practice is cumbersome, largely because quantifying PC density or morphology pattern is time-consuming and expertise-dependent, and thus not suitable for daily practice. Therefore, developing an automated, unsupervised algorithm that can quantify PC density would allow us to efficiently use PC as a biomarker to enhance clinical evaluation and care for patients with CD.

Recently, deep learning (DL) methods have been transformational in developing robust and accurate tools for complex and time-consuming image evaluation challenges such as PC quantification.[22] DL techniques, particularly convolutional neural networks (CNNs), are widely used on large datasets of images such as tumour detection, segmentation of organs, abnormality detection, and classification within medical images.[23,24] CNNs are a powerful tool due to their ability to learn complex patterns from large-scale medical imaging data, their robustness to variability, and their state-of-the-art performance for a wide range of medical image analysis tasks. We have previously used CNNs to generate algorithms to quantify the extent of steatosis in the context of donor liver examination,[25] as well as glomerulosclerosis for donor kidney evaluation.[26,27]

Algorithms that may have clinical relevance to evaluate histopathologic features of IBD are being developed, such as for mucus (a product of goblet cells)[28] and PC granule areas,[22] and for prediction of outcome in paediatric ulcerative colitis[29] as reported. While these are of interest, none of these algorithms provide accurate goblet cell or PC densities as readouts, thus the applicability of these algorithms is limited. Herein we report the development of a DL algorithm for PC density quantification and demonstrate that this algorithm is capable of accurately determining PC densities on histopathology slides and is comparable to assessments made by expert gastrointestinal (GI) pathologists. We further confirmed the real-world applicability of this algorithm PC density differences using whole slide images (WSI) from a large clinical cohort, showing that PC density correlates with disease status and risk of recurrence in CD. Thus, the PC algorithm could be an ancillary tool to improve the care for patients with IBD.

Methods

Ethics

This study was performed in compliance with the approved Institutional Review Board protocol #201703119 at Washington University School of Medicine. This study is classified as a retrospective/secondary data analysis study and was approved for waiver of consent by the institutional review board.

Sex and gender reporting

Tissue from male and female sexes were used for the study. There was no selection or exclusion based on sex or gender for the current study.

Patient samples

Haematoxylin & eosin (H&E)-stained histopathology sections cut from formalin-fixed paraffin embedded tissue blocks from archival clinical cases were used for secondary analysis for this study.

The sections used for algorithm training included cases identified from a small series of consecutive histologically normal terminal ileum biopsies and ileal resection margins identified by study pathologists in the course of clinical service, as well as ileal resection margin specimens from ileocolectomy cases from patients with IBD that had been curated for another clinical retrospective study, spanning from 2009 to 2013. No demographic or other selection criteria were used for the training slide selection. These cases are summarised in Supplementary Table S1.

The validation study patient cohort, retrospective by design, consisted of a total of 190 slides from Barnes-Jewish Hospital in St. Louis, Missouri, USA and Cedars-Sinai Medical Center, Los Angeles, California, USA, a subset of which were previously described in a cohort study of patients with CD,[15] as part of IBD consortium studies funded by the Helmsley foundation SHARE consortium and NIDDK IBD Genetics Consortium. Briefly, patients with CD were prospectively recruited between 2005 and 2013 at Washington University/Barnes-Jewish Hospital and between 1999 and 2013 from Cedars-Sinai. Clinical metadata of the patients collected include demographics, family history, disease duration, CD clinical phenotype (per Paris criteria), and follow up information (postoperative prophylaxis, disease recurrence) after surgery. Disease recurrence was defined by endoscopy and/or radiology as we previously described.[15–17] For follow-up time interval, the start time was defined as the date of surgery, and the end date was defined as the date of confirmed recurrence, or if no recurrence occurred by the last known follow up, the patient was censored at the date of last known clinical follow-up. Additional ileal resection specimens from patients without IBD in the from the same consortium enrolment (2006–2013), were used as cross-sectional study controls for one of the experiments. All non-IBD cases were from Barnes-Jewish Hospital as previously reported.[18] No matching criteria were used to select patients without IBD patient samples. All available slides from Barnes-Jewish Hospital and Cedars-Sinai cohorts were scanned (see below for technical details of scanning) and acceptable quality scans were used for the validation study.

Whole slide imaging

H&E-stained histopathology sections were scanned for WSI. Slide scanning was performed using Hamamatsu NanoZoomer 2.0-HT System (Alafi Neuroimaging Laboratory, the Hope Center for Neurological Disorders) or Aperio XT (Digital Imaging Center, Department of Pathology & Immunology) scanners at 20× and 40× objective magnification, respectively. The initial round of annotations was conducted using 40× images, while all subsequent data was collected at 20X. To ensure consistency, the initial round of annotations was resolved to 20×. Each slide had 0.5 microns per pixel (mpp).

Image annotation by GI pathologists

Study pathologists used a custom graphical user interface, written as a plugin for ImageJ, to manually create the annotations.[30] Due to ImageJ memory constraints, each WSI was sliced into sub-images of approximately 10,000 pixels on a side. Each crypt was denoted by an outline drawn around it while each PC was denoted by point annotation (Fig. 1a and b). For each input image, 3 separate pixel-mapped targets with the same spatial dimensions from the annotations were generated. The 'crypt-label' target sets pixels '1' in areas associated with crypts containing PCs and zero elsewhere; the 'PC-label' target sets pixels '1' in all areas within a 10-pixel radius of each PC and zero elsewhere; and the 'PC-number' target creates a probability mapping across each crypt area with the values corresponding to the total number of PCs included within that area of the crypt (zeros elsewhere). The resulting training dataset of 122 sub-images ("patches"), from 14 annotated WSI, identified a total of 4338 crypts with PCs and 20,065 PCs. The clinicopathologic demographic information for the training set is provided in Supplemental Table S1.

DL model architecture

Fig. 1c depicts a graphical overview of the model training and prediction procedure. Briefly, we applied a two-stage approach. The stage 1 model uses image patches extracted from the training WSI dataset to train against patches extracted from the corresponding targets. WSIs were processed by the algorithm tile by tile — with appropriate padding — and reassembled into a heatmap covering the entire slide. We used conventional segmentation techniques to threshold the heatmap and identify segment crypt regions from predictions. The number of PCs in each crypt was either read directly from the 'PC-number' output or input into another model (stage 2) to further refine the computation of the number of PCs in each crypt. We used Tensorflow version 2.13.1 to code the model and 1.16 GB T V100 GPU to train and test the model.

Stage 1 model

Fig. 2a depicts the stage 1 model, a modified U-net architecture[31] with three outputs. The outputs corresponded to the three targets: 'crypt-label', 'PC-label', and 'PC-number'. The 'crypt-label' and 'PC-label' outputs minimise cross-entropy loss, while the 'PC-number' output minimises the sum-of-squares loss. We weighted the 'crypt-label' and 'PC-label' losses at each pixel 20:1 and 10:1 relative to the background, respectively, to account for class imbalance. The model was trained with the Adam optimiser (learning rate of 10^{-4}) for 30 epochs

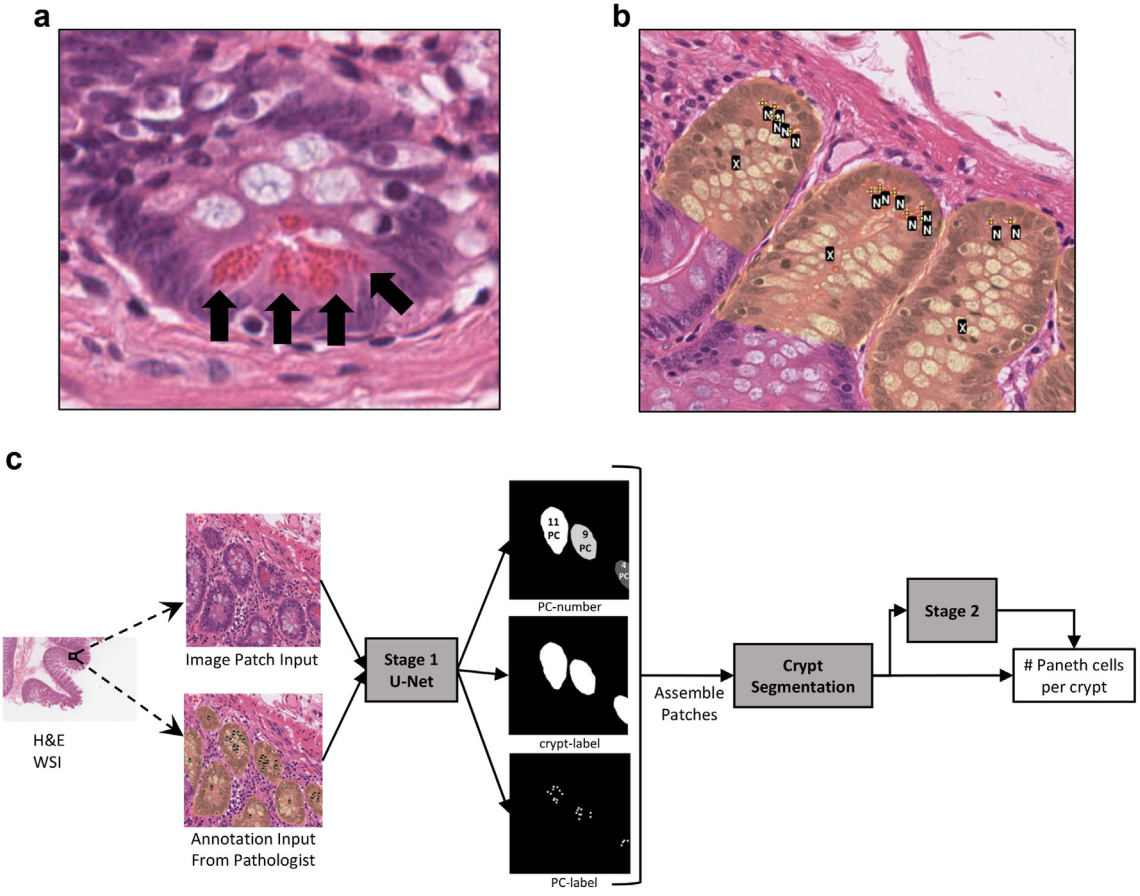

***Fig. 1:* Overview of quantification of PCs using a deep learning algorithm.** a) PCs (marked by arrows) are shown in an intestinal crypt on H&E stain. Each whole slide image (WSI) of H&E-stained histology slides is processed into sub-images for annotation b) by the study pathologists, with annotation mask denoting crypts in "X" and individual PCs in "N". c) The Stage 1 model trains on image patches of these WSI, along with input annotations from study pathologists denoting the crypt and PCs present in the image. The Stage 1 model utilises a U-net architecture. Stage 2 uses the output from Stage 1 to improve the accuracy of the number of PCs per crypt.

using TensorFlow in minibatches of 8. To generate the training input, we divided each training image into overlapping image patches (892 × 892 pixels, stride of 678 pixels) and only kept those that contained at least 20% non-whitespace. No preprocessing was performed on the input data. Image augmentation was performed with random 0/90/180/270-degree rotation and up/down flipping. The training was performed with 5-fold cross-validation, such that 5 models were each trained with-holding a separate 20% of the complete dataset, and evaluation was performed on the withheld data.

The stage 1 model 'crypt-label' output helps to determine the position and spatial extent of crypts with PCs. After stitching stage 1 prediction patches into an image co-registered with the input H&E image, a threshold level of 0.8 on the 'crypt-label' result yields connected candidate crypt regions. These were individually segmented and labelled using the open-source scikit-image package.[32] A threshold value of 0.5 on the stitched 'PC-label' indicates the location of PC areas. Any region without at least one corresponding positively labelled 'crypt-label' pixel within that region's boundary was excluded. The coordinates of the resulting regions segment the stage 1 model 'PC-number' output. In the initial version of the study protocol, the number of PCs in a specific crypt region was read directly from the maximum value of the 'PC-number' output within that region.

Stage 2 model

To improve the final PC counting performance, a second stage of modelling was performed. This model used the output of the stage 1 model as input, specifically the 'PC-label' output segmented by crypt. Fig. 2b depicts the stage 2 model which uses a fully convolutional model architecture. The model output received 280 × 280 pixel patches as input and output a single number giving the number of PCs in the patch. The stage 1 'PC-label' SoftMax prediction provided the patches for input, where

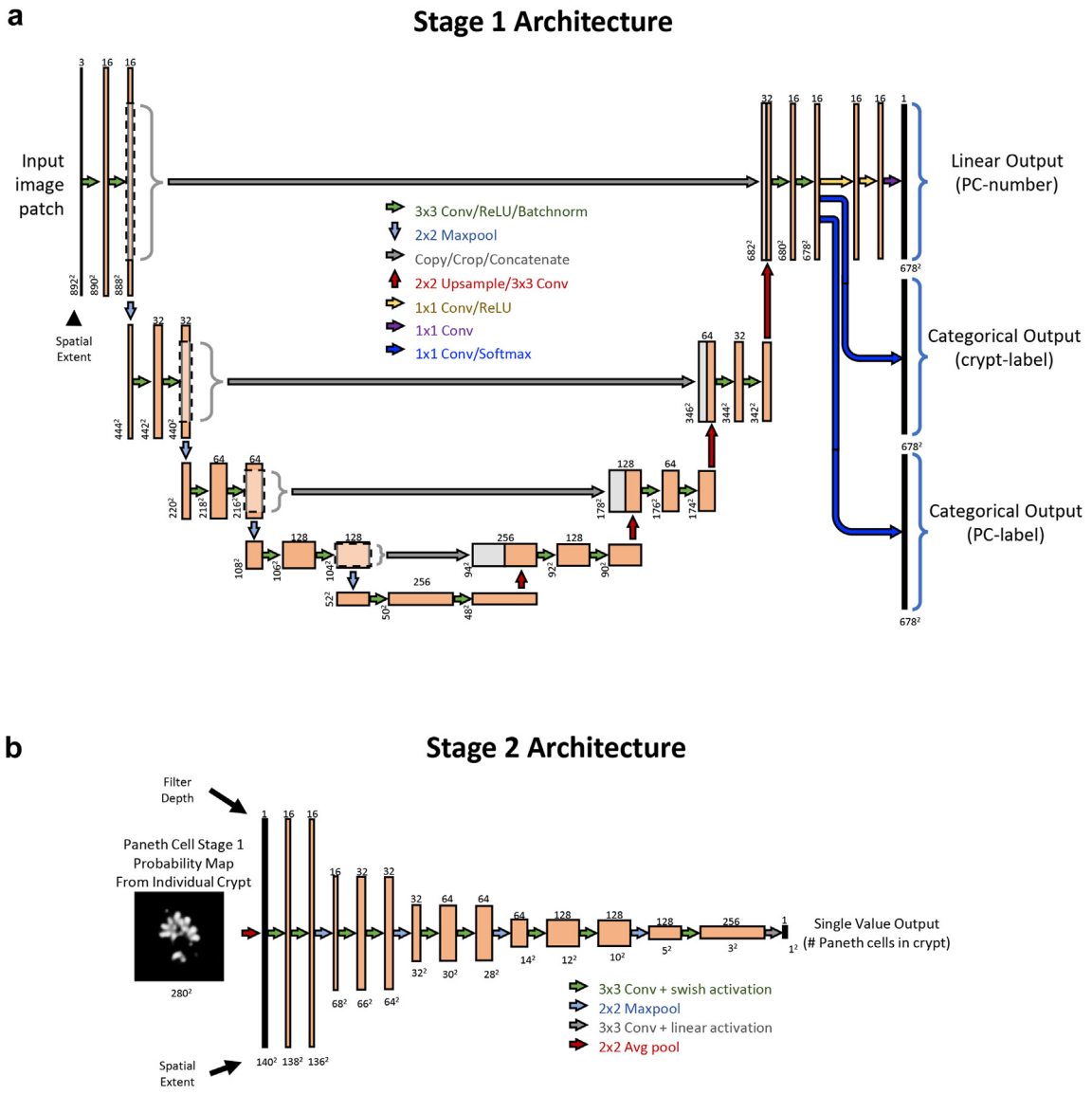

***Fig. 2:* Detailed graphical overview of deep learning model architecture.** a) The Stage 1 model trains on image patches of the whole slide image using a U-net architecture with three output targets. b) Stage 2 uses the output from Stage 1 as input, and the output from Stage 2 is a single value of PC density (number of PCs per crypt).

each patch was centred on a detected crypt. The target value was the number of PCs given in the associated crypt annotation. Input data was augmented with random 0/90/180/270-degree rotation and up/down flipping. The model was trained with 5-fold cross-validation to minimise the sum-of-squares loss between the output and target value using the Adam optimiser with a learning rate of 10^{-6} for 200 epochs in mini-batches of 8.

Validation studies

For an external validation of the model, a set of 48 new WSIs were collected. We chose a single region of interest (ROI) for each of the new WSIs covering, on average, 148 crypts per ROI. Each ROI was annotated by at least two GI pathologists counting the number of PCs and the number of crypts using QuPath.[33] The average number of PCs per crypt was computed from these numbers. The fully trained stage 1 and 2 models were used to predict these new ROIs.

In addition to the validation study, another 142 WSIs (with both patients with CD and patients without IBD included) were analysed with the stage 1 and 2 models, and PC/crypt was calculated. Clinical demographics and outcomes were extracted from any available clinical

records for all WSIs. Portion of this cohort was previously described.[15]

To compare the time it takes for the DL model to analysed and calculate the PC density in comparison to manual counting of crypts and PCs by pathologists, ten images were randomly selected from the training set images using GraphPad free Random Number Generator (https://www.graphpad.com/quickcalcs/randomN1/). These images were evaluated by two study GI pathologists, and time needed to finish counting was recorded. The same images were analysed by the DL model and processing time was recorded.

Sample size estimation

The number of WSI needed to design and optimise the Paneth cell DL algorithm was estimated from our previous work developing DL algorithms for liver steatosis and glomerulosclerosis,[25–27] in which 8458 glomeruli or 10,653 annotations for foci of hepatic steatosis were used to train the respective models. In the 122 sub-images we generated from small intestinal 14 WSI, there are annotations for 4338 crypts with PCs and 20,065 PCs. Given the similar scale of annotated features of interest in the training set, we predicted this would be adequate based on our previous experiences with the other CNN models. The number of CD samples for outcome correlation was based on previous studies from our group and others in examining Paneth cell morphology phenotype and time to recurrence after surgery in CD,[15–17,21] which was previously established as requiring at least 90 patients in a cohort study to resolve differences in recurrence-free survival.[17]

Statistics

The stage 1 model uses the whole slide images as the independent variable. The dependent variables are a pixel-level map of (1) whether a pixel falls within a crypt, (2) false within a PC, and for crypt-overlapping pixels, (3) the number of PC in this specific crypt, as described in the DL Model Architecture section (Fig. 2a).

For the stage 2 model, the independent variables are the outputs of stage 1. The dependent variable is the number of PCs within each crypt, assessed at a per-crypt level, as described in DL Model Architecture section (Fig. 2b).

Several metrics were used to evaluate performance including root mean square error (RMSE) and the coefficient of determination (r^2). RMSE measures the average difference between the model predictions and the ground truth. A lower RMSE is optimal as this implies that the predictions are close to the ground truth. The coefficient of determination is a measure of how well a model fits the data. A high r^2 implies there is a strong correlation. We calculated these metrics for the average number of PCs per crypt, the total number of crypts, and the total number of PCs. Non-parametric comparison between two groups was performed using the Mann–Whitney test. Survival curves were compared using Log-rank (Mantel–Cox) test. Statistics were performed using the Python library sklearn[34] and GraphPad Prism v10 (GraphPad Software, Boston, Massachusetts USA, www.graphpad.com).

Sensitivity and specificity at a pixel-level on the crypts and PC masks were used to evaluate the stage 1 performance of the model. Sensitivity, or the true positive rate, refers to the proportion of positive pixels (labelled with a 1) with a prediction greater than 0.5. Specificity, or the true negative rate, refers to the proportion of negative pixels (labelled with a 0) with a prediction less than 0.5, though the large regions of easy to classify negative pixels may confound this metric. These metrics were computed using all the pixels in each slide alone, and then averaged across all slides.

Role of funders

The funder (NIH) did not have any role in study design, data collection, data analyses, interpretation, or writing of report or submission decision.

Results

Patient characteristics

We scanned histology slides of ileal resection margins ileal/ileocolonic resections or endoscopic biopsies from patients with or without IBD to generate WSIs for generation of DL model. The characteristics of the patients whose ileal resection or slides were utilised for the training set are listed in Supplementary Table S1. In the training set, eight out of 14 patients (57.1%) of patients were male, with a mean age of 47.2 years at time of procedure. Eleven of 14 patients had a history of IBD (78.6%); however, the majority of cases had non-inflamed terminal ileum in the histologic sections used (85.7%).

For DL algorithm application in CD and non-IBD patients ("validation set"), we obtained WSIs from uninflamed ileal sections from additional 48 non-IBD and 142 CD patients (Table 1). The mean age of the patients at the time of surgery was 62.1 years in the non-IBD cohort and 38.6 years in the CD cohort. In the non-IBD cohort, 43.75% of the patients were male, while 49.3% of the patients with CD were male. The average BMI at time of surgery was 30.7 in non-IBD patients, while patients with CD had an average BMI of 23.0. Both cohorts were predominantly White/Caucasian (79.17% in non-IBD patients, 91.55% in CD patients). Multiple disease-specific clinical parameters were also extracted for the CD cohort. The average age of diagnosis with CD was 27 years of age, with an average interval between the time of diagnosis and the present surgery of 10.92 years. The majority of patients with CD had either ileal disease (49.30%) or ileocolonic disease (38.73%). 23.24% also had perianal disease. Most had stricturing disease (67.61%), and 40.14% had penetrating disease behaviour. 76.06% had post-operative

prophylactic treatment. The mean length of follow up was 16.78 months (n = 138; SD ± 23.10 months).

Output visualization of model-predicted crypts and PCs

Graphs were generated to visualise the output of the model in comparison to the annotated "ground truth" training images, in an *in situ* spatial orientation, for both crypt identification, as well as predicted number of PCs for each crypt. The Stage 1 predicted image outputs were qualitatively similar to the target annotation maps generated from the pathologist annotations (Fig. 3). The annotations given included where the crypt was and the exact number of Paneth cells per crypt. Stage 1 predicted where each crypt was located, shown by an outline. Stage 1 also predicted a linear output, or probability mapping, over each crypt. When this data was displayed, the colour mapping reflects where within each crypt the model predicted Paneth cells as opposed to a total number per crypt like the annotations.

Evaluation of stage 1 and 2 models

The stage 1 model performed well when modelling the total number of crypts (RMSE = 6.872, r^2 = 0.962) and the total number of PCs (RMSE = 78.230, r^2 = 0.913) (Fig. 4a). However, the stage 1 model consistently predicted the average number of PCs per crypt slightly higher than the ground truth, resulting in lower performance (RMSE = 1.880, r^2 = 0.641). When we incorporated the use of stage 1 and 2 models for predictions, this improved the performance of the average number of PCs per crypt (RMSE = 0.802, r^2 = 0.748; Fig. 4b). Modelling performance for the total number of crypts remained consistent (RMSE = 6.872, r^2 = 0.962), while the performance for the total number of PCs improved slightly but not significantly (RMSE = 51.736, r^2 = 0.919).

We then computed the same metrics using the max and the mean of the predictions from Stage 1 as a baseline comparison to determine whether the stage 2 model was necessary. There was a decrease in performance using the max (RMSE = 1.880, r^2 = 0.641) and mean (RMSE = 1.830, r^2 = 0.636) of the predictions from Stage 1 alone (Supplementary Figure S1). Thus, the addition of the stage 2 model was also an improvement over baseline metrics.

The model demonstrated strong performance in identifying crypts and PCs. The model correctly identified 75.1% of crypts pixels. The specificity for crypt pixels was 99.6%, reflecting a favorably high true positive rate. The model also successfully identified 87.8% of PC pixels. The specificity for PC pixels was 99.9%, indicating that the model accurately identifies true negatives with virtually no false positives.

Comparison of DL model to expert GI pathologists' annotations

Using a subset of the additional validation WSIs, study pathologists manually quantified crypts and PC to

	Non-IBD controls	Crohn's Disease cohort
	n = 48	n = 142
Paneth cell per crypt (Stage 1 + 2; ±SD)	4.04 (±0.71)	2.99 (±1.12)
Age at surgery, mean age in years (n; ±SD)	62.1 (n = 48; ±13.3)	38.6 (n = 140; ±16.0)
Age at diagnosis, mean age in years (n; ±SD)	N/A	27 (n = 135; 14.34)
Sex, n (%)		
Male	21 (43.75)	70 (49.3)
Female	27 (56.25)	71 (50)
Data not available	0 (0)	1 (0.70)
Body Mass Index, mean (n; ±SD)	30.7 (n = 44; ±9.4)	23.0 (n = 126; ±4.90)
Race, n (%)		
Caucasian	38 (79.17)	130 (91.55)
African-American	9 (18.75)	8 (5.63)
Asian	1 (2.08)	2 (1.41)
Hispanic	0 (0)	0 (0)
Other	0 (0)	1 (0.70)
Data not available	0 (0)	1 (0.70)
Smoking history (n, %)		
Never smoker	22 (45.83)	90 (63.38)
Active or ex-smoker	14 (29.17)	47 (33.10)
Data not available	12 (25.00)	5 (3.52)
Duration between diagnosis and surgery, mean age in years (n; ±SD)	n/a	10.92 (n = 134; ±10.71)
Family history of IBD, n (%)		
No	n/a	54 (38.03)
Yes	n/a	22 (15.49)
Data not available	n/a	66 (46.48)
Disease location (Paris criteria), n (%)		
L1: distal ileal	n/a	70 (49.30)
L2: colonic	n/a	4 (2.82)
L3: ileocolonic	n/a	55 (38.73)
L4: upper GI involvement	n/a	11 (7.75)
Data not available	n/a	2 (1.41)
Perianal disease, n (%)		
No	n/a	95 (66.90)
Yes	n/a	33 (23.24)
Data not available	n/a	14 (9.86)
Disease behaviour: stricturing, n (%)		
No	n/a	31 (21.83)
Yes	n/a	96 (67.61)
Data not available	n/a	15 (10.56)
Disease behaviour: penetrating, n (%)		
No	n/a	69 (48.59)
Yes	n/a	57 (40.14)
Data not available	n/a	16 (11.27)
Post-operative prophylaxis (summary), n (%)		
No	n/a	25 (17.61)
Yes	n/a	108 (76.06)
Data not available	n/a	9 (6.34)
Post-operative prophylaxis: Immunomodulators, n (%)		
No	n/a	52 (36.62)
Yes	n/a	81 (4.91)
Data not available	n/a	9 (6.34)

(Table 1 continues on next page)

	Non-IBD controls	Crohn's Disease cohort
	n = 48	n = 142
(Continued from previous page)		
Duration of follow-up until recurrence or last follow-up, mean in months (n; ±SD)	n/a	16.78 (n = 138; ±23.10)

Table 1: **Clinical demographics and parameters for whole slide image clinical cohort.**

calculate PC density. To demonstrate the rigor of the algorithm, the performance of the validation study was compared to the manual annotation of the senior expert GI study pathologist, with decades of expertise in PC manual quantitation[15–19] (RMSE = 1.148, r^2 = 0.708; Fig. 5a) and was very similar to the performance of the average number of PCs per crypt based on the training dataset with cross-validation (Fig. 4b). The other two GI fellowship-trained pathologists also had comparably good correlation with the algorithm output (Fig. 5b). This result demonstrates the generalizability of the stage 1 and 2 model system.

To evaluate for the efficiency of the algorithm compared to the time it would take for a pathologist to manually quantify numbers of crypts and PCs, the time needed to tabulate the number of crypts and PCs in an image was compared to manual quantification by two study pathologists, portrayed graphically in Supplementary Figure S2. The algorithm took an average of 20.598 s (range 2.3–40.8 s) to process each image and generate a result. For the same images, it took Pathologist 1 an average of 227.4 s (range 75–475 s) and Pathologist 2 an average of 398.2 s (range 103–887 s). The algorithm was faster than both pathologists for all ten cases. The time efficiency of the algorithm compared to the pathologist time spent ranged from a factor ~2.5 times more efficient to over 180 times.

DL algorithm-generated PC density correlates with clinical outcome in CD

As a proof-of-concept experiment, we next defined the PC densities in the ileal section slides from all of the validation CD and non-IBD cases using the DL algorithm. The average PC density in patients with CD was significantly lower than that of the patients without IBD (4.04 [±0.71]) cells/crypt for patients without IBD versus 2.99 (±1.12) cells/crypt in patients with CD ($p < 0.0001$) (Fig. 6a), which is consistent with what was reported in the literature.[21,35–37] PC density did not appear to correlate with patient demographics (Supplementary Table S2). Using the 25 percentile (first quartile) of the average PC density among patients with CD as a cut-off for PC density "high" versus "low", we found that patients with low PC density were more likely to have an early postoperative recurrence ($p = 0.0399$) (Fig. 6b). Therefore, the PC density DL algorithm may be used to predict outcomes in patients with CD undergoing surgical resection.

Discussion

Dysfunction and loss of PCs have been well-established as risk factors and part of the disease progression in a number of diseases affecting the small intestine, including graft-versus-host-disease[38] and IBD, particularly CD.[15,39,40] Clinically validated biomarkers to predict recurrence in CD are lacking, and while numerous studies have quantified PC number and morphology in human patients and found a strong correlation to the outcome,[17,21] these were time-intensive manual counts of up to hundreds of crypts and frequently utilise immunohistochemical stains to highlight the PCs for counting. This process is impractical from both a time and resource standpoint in the clinical setting. Thus, the goal of this study was to use DL to develop a robust way of accurately quantifying the number of PCs in the small intestine on H&E-stained slides. To date, to our knowledge, there has been no other study that has attempted to accurately quantify PC cell number and density using computational pathology for a possible future application as a biomarker for CD prognosis and treatment response, among other possible applications.

The DL algorithm we have developed successfully demonstrates that it is feasible to use DL models to accurately count PCs and crypts in a WSI. While we initially achieved strong correlation to the ground truth with the Stage 1 quantification of crypts and PCs separately, the final density was less accurate. However, refinement of the model with a second stage greatly strengthened the correlation of the algorithm output to the training set. Moving forward with this two-stage model, we were able to demonstrate translatability of the model to a large validation cohort of 190 WSI. The performance of the model in analysing PC density in the validation cohort WSIs was comparable to its performance in cross-validating to the training set, with an expert GI pathologist providing the ground truth in the validation cohort for comparison. Key confirmation of the powerful potential clinical utility of the DL algorithm is demonstrated in the clinical correlation studies we performed on the validation cohort: using PC density as a single quantitative histologic biomarker, we were able to confirm that subjects with CD harbour fewer PCs/crypt compared to subjects without IBD, and more importantly, showed that the lowest quartile of PC density among patients with CD correlated with a higher risk of recurrence.

The development of this model is pivotal for using PCs as a biomarker because this type of quantitative assessment would not otherwise be possible to integrate broadly into healthcare practice. The DL model can quantify PCs more efficiently, with the current model able to process an image anywhere from ~2.5 times to

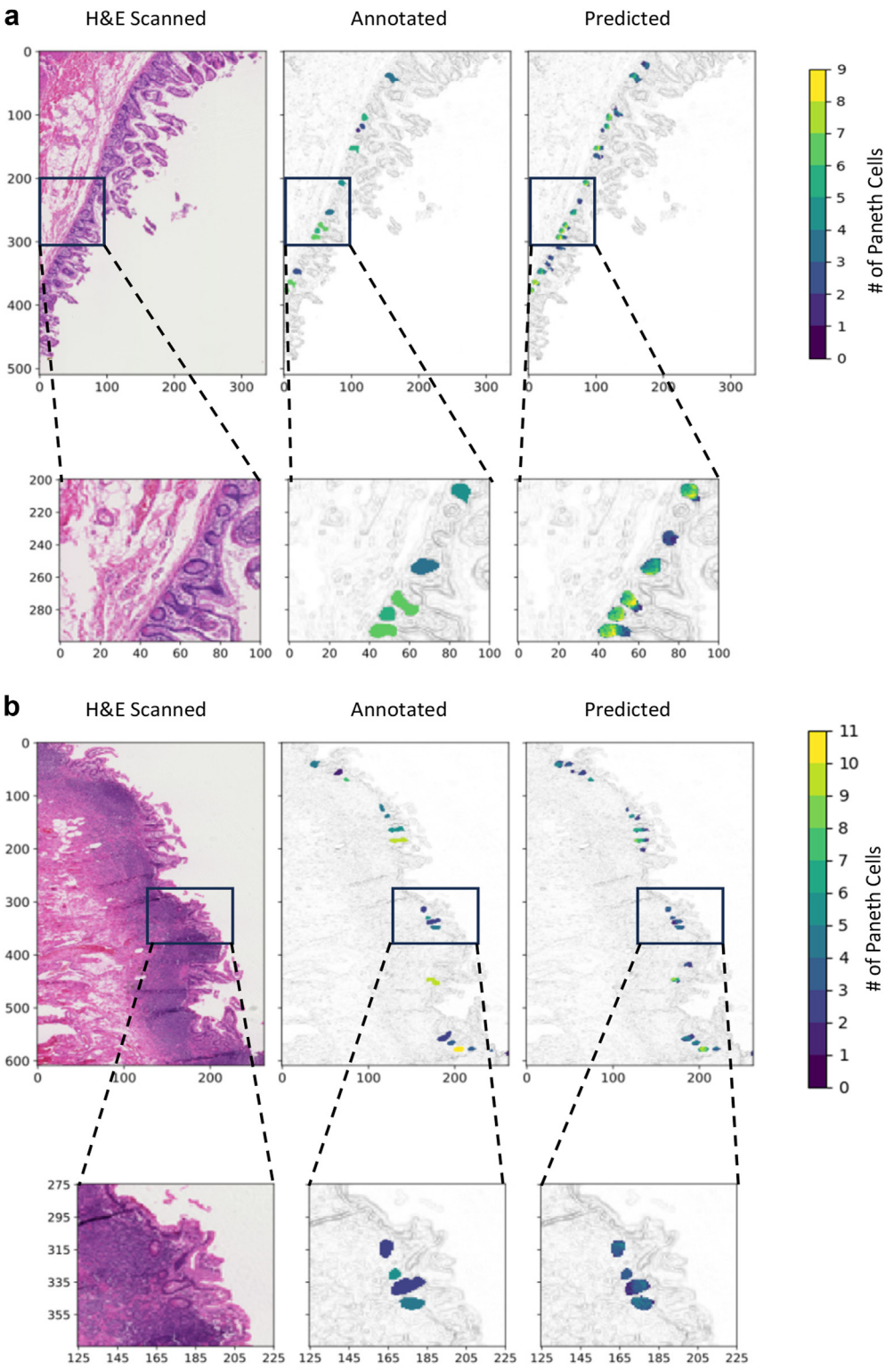

Fig. 3: **Stage 1 predictions are qualitatively similar to the target annotation maps.** a and b) Representative WSIs used for training, along with annotation and predicted PC numbers. The H&E scanned image (left) was annotated by a study pathologist (centre). The annotations were used as targets during training for Stage 1. The Stage 1 predictions are displayed on the right panels. The pathologist annotation and Stage 1 predictions are depicted with tissue overlay for orientation. Additionally, corresponding selected areas at higher magnification are shown in the outset plots. Axes scales are XY coordinates by pixel number.

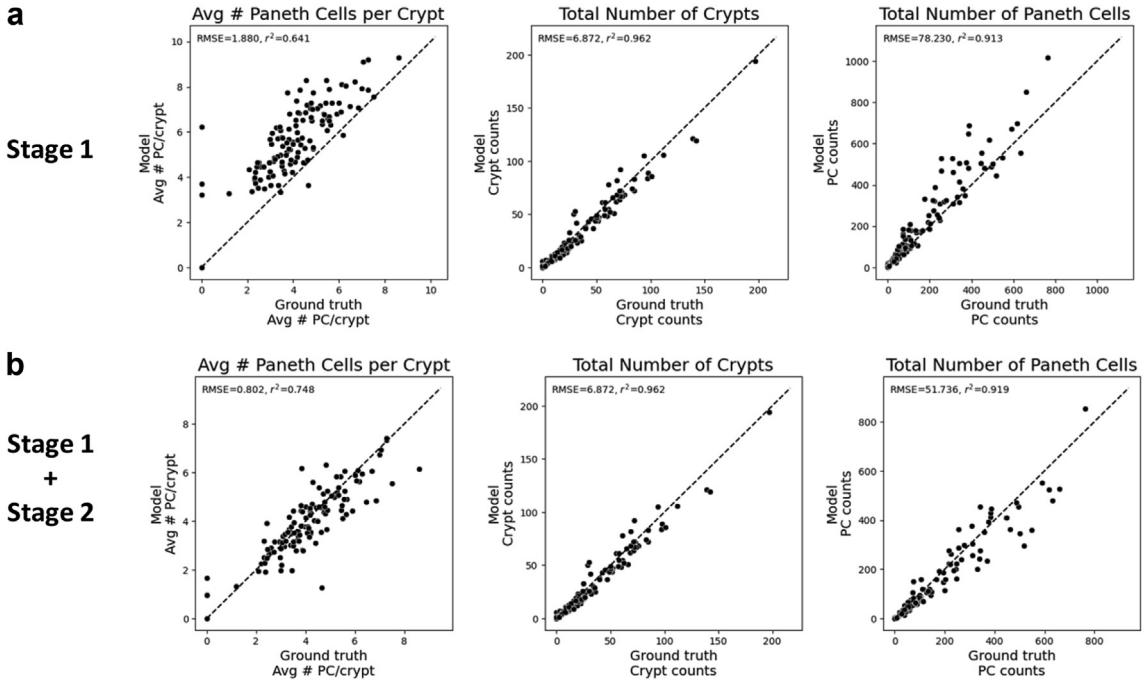

Fig. 4: **Modelling crypts and PCs is more accurate with Stage 1 + Stage 2 in cross-validation of training results.** a) Stage 1 model performs well when modelling the total number of crypts and the total number of PCs (*RMSE* = 6.872, r^2 = 0.962, *RMSE* = 78.230, r^2 = 0.913, respectively). b) When stage 2 is incorporated, performance of the average number of PCs per crypt improves (*RMSE* = 0.802, r^2 = 0.748).

>150 times faster than pathologists. Additionally, the algorithm can run in the background on a computer or a server. But beyond absolute time savings of a few or several minutes for the pathologist, the conservation of the intense mental energy that it takes to efficiently count images for quantitative biomarkers cannot be overstated from the perspective of the pathologist. Rescue from the cumulative mental fatigue of quantitative biomarker manual evaluation frees up the pathologist to focus their concentration on more complex diagnostic challenges.

While this study demonstrates the feasibility of quantifying Paneth cells using the proposed algorithm, it is important to acknowledge that it serves as a proof of

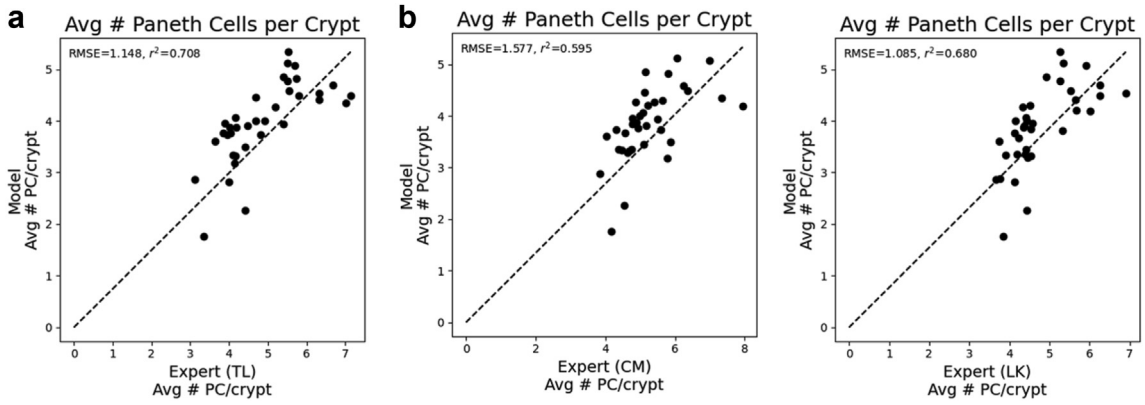

Fig. 5: **Stage 1 + Stage 2 algorithm correlates similarly well with pathologist annotations in validation study as it did in model cross-validation.** Study pathologists annotated a new WSI validation cohort. a) The model performed closely with the pathologist's annotations (representative pathologist: TL: *RMSE* = 1.148, r^2 = 0.708). The performance of the model is comparable to the cross-validation study performance shown in Fig. 4B (*RMSE* = 0.802, r^2 = 0.748). b) The algorithm also compared well with the density annotations of two other GI-fellowship trained pathologists.

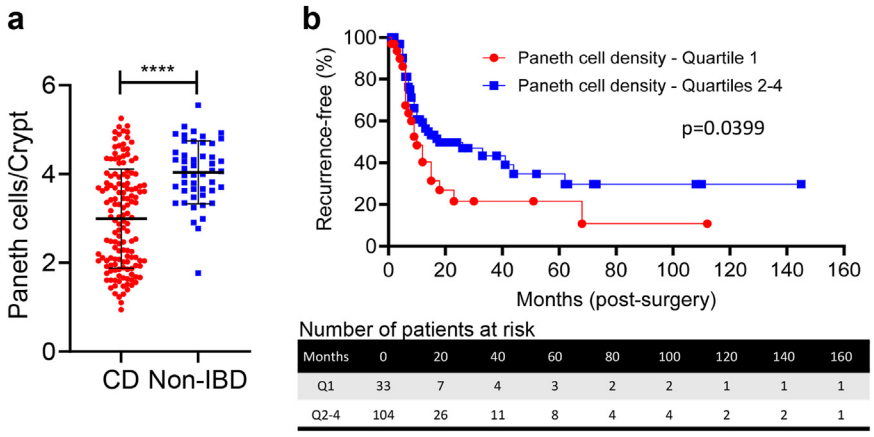

***Fig. 6:* Paneth cell density, as calculated by the deep learning algorithm, strongly correlates with disease status and recurrence-free survival.** a) Patients without IBD (n = 48; mean of 4.04 [±0.71]) cells/crypt) have significantly higher PC density than patients with CD (n = 142; mean 2.99 (±1.12) cells/crypt), p < 0.0001 by Mann–Whitney test; bars shown are mean ± SD. b) Patients with CD with the lowest 25% PC density (Quartile 1) have significantly shorter recurrence-free interval (p = 0.0399; Log-rank test).

Plain language summary

In gastrointestinal (GI) inflammatory diseases, like CD, changes in number and quality of PCs, specialised intestinal cells that help with host immunity and gut health, can be associated with disease outcomes. However, counting these cells by hand is very time-consuming, making it difficult for doctors to efficiently quantify PCs as a predictive marker for use in clinical practice. To address this, we developed a computational model to quickly and accurately count PC density. We trained the model using tissue samples from people with and without inflammatory bowel disease, for which pathologists identified all the PCs ahead of time. After testing the trained model, we added a second stage model, which takes the output from Stage 1 and improves accuracy of predicted PC density. Fig. 7 shows a flow diagram of the inputs and outputs of the model. We compared the results of the Stage 2 model to counts provided by experienced pathologists on a validation set of images and found the two-stage model was highly accurate in quantifying PC density. Using the model, we found that intestinal samples from patients without IBD had a higher density of PCs than patients with CD. Furthermore, when we used the model to study samples from people with CD, we found that those with lower density of PCs had shorter time to disease recurrence. This suggests that the number of PCs could help predict how CD will progress. In the future, this tool could help doctors more accurately predict risk of recurrence for a patient with CD, among other applications. This study demonstrates feasibility for measuring PCs to help predict disease outcomes in clinical settings.

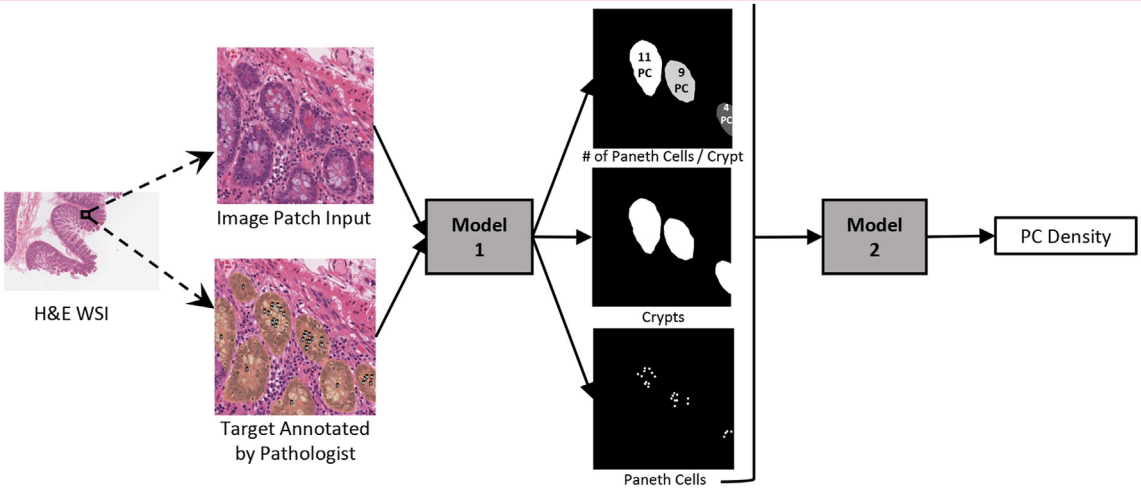

***Fig. 7:* Overview of quantification of PCs using a deep learning algorithm**. The stage 1 model trains on image patches of these WSI, along with input annotations from study pathologists denoting the crypt and PCs present in the image. Stage 2 uses the output from Stage 1 to improve the accuracy of the number of PCs per crypt.

concept rather than a fully deployable clinical tool. There are some limitations to this study and additional validations to be done before this prototype can be further developed for possible clinical application. This study used a modestly sized dataset for training and validation, where all WSIs were scanned by the same two machines and were of similar quality. To be ready for direct and broad use, the training and validation datasets should be expanded to include WSIs from a range of scanners with different scanning qualities. In addition, there is also lack of stain normalization, which may impact the model's robustness across varying staining conditions commonly encountered in clinical practice. In future work, we would likely benchmark stain normalization and other techniques to manage stain variability, which could further enhance the algorithm's reliability and generalizability. Further validation with more CD cohorts with accompanying clinical outcomes, and a prospective study to see if the algorithm can predict recurrence with comparable or improved accuracy to other clinical indicators, are also needed.

A natural extension of this work would be to leverage DL to not only quantify PCs, but also identify abnormal PC morphologies and quantify the relative proportion of normal to abnormal PCs, the ratio of which has been demonstrated by our group and others to correlate with patient outcome in CD.[15,17] In addition, we recently showed that PCs also correlate with clinical outcomes in patients with ulcerative colitis (another major form of IBD).[41] Therefore, our DL algorithm can potentially be used in all patients with IBD histopathology evaluations.

In summary, we developed a two-stage DL model that successfully demonstrates feasibility, generalizability, and efficiency of use. Further development of this algorithm may aid in facilitating the integration of PC density as a biomarker into clinical practice for intestinal diseases.

Contributors

The authors confirm their contribution to the paper as follows: Study conception and design: TCL, SJS, TSS; data collection: LIK, CM, LL, PK, JNM, TCL, TH, EM, DPBM, AM; analysis and interpretation of results: KS, LIK, TCL, JNM, SJS; draft manuscript preparation: KS, LIK, JNM, TCL. All authors have been given free access to results, approve for publication, and accept associated responsibility. All authors reviewed the results and approved the final version of the manuscript.

Data sharing statement

The deidentified datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declaration of interests

The authors have no conflict of interest to declare in relation to the work presented.

Acknowledgements

We thank the Alafi Neuroimaging Laboratory, the Hope Center for Neurological Disorders, and NIH Shared Instrumentation Grant (NCRR 1S10RR027552) to Washington University for core facility resources used for generation of data for this manuscript. We also thank the Digestive Disease Research Core Center at Washington University (supported by P30 DK052574) and the Cedars-Sinai MIRIAD IBD Biobank.

Research reported in this publication was supported by the National Institutes of Health (R01 DK124274, DK125296, DK136829, and DK138465 to TCL). LIK was supported by T32 EB021955 and K12 CA167540. SJS was supported by NIH R01DK12427. KS was supported by R01 DK124274. DPBM was funded by NIH/NIDDK U01 DK062413. DPBM, TSS, and TCL were additionally funded by The Leona M and Harry B Helmsley Charitable Trust.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.ebiom.2024.105440.

References

1 Bevins CL, Salzman NH. Paneth cells, antimicrobial peptides and maintenance of intestinal homeostasis. *Nat Rev Microbiol*. 2011;9(5):356–368.
2 Adolph TE, Mayr L, Grabherr F, Tilg H. Paneth cells and their antimicrobials in intestinal immunity. *Curr Pharm Des*. 2018;24(10):1121–1129.
3 Stappenbeck TS. Paneth cell development, differentiation, and function: new molecular cues. *Gastroenterology*. 2009;137(1):30–33.
4 Clevers HC, Bevins CL. Paneth cells: maestros of the small intestinal crypts. *Annu Rev Physiol*. 2013;75:289–311.
5 Wallaeys C, Garcia-Gonzalez N, Libert C. Paneth cells as the cornerstones of intestinal and organismal health: a primer. *EMBO Mol Med*. 2023;15(2):e16427.
6 Sato T, van Es JH, Snippert HJ, et al. Paneth cells constitute the niche for Lgr5 stem cells in intestinal crypts. *Nature*. 2011;469(7330):415–418.
7 Pentinmikko N, Iqbal S, Mana M, et al. Notum produced by Paneth cells attenuates regeneration of aged intestinal epithelium. *Nature*. 2019;571(7765):398–402.
8 Armbruster NS, Stange EF, Wehkamp J. In the Wnt of Paneth cells: immune-epithelial crosstalk in small intestinal Crohn's disease. *Front Immunol*. 2017;8:1204.
9 Yu S, Balasubramanian I, Laubitz D, et al. Paneth cell-derived lysozyme defines the composition of mucolytic microbiota and the inflammatory tone of the intestine. *Immunity*. 2020;53(2):398–416.e398.
10 Courth LF, Ostaff MJ, Mailander-Sanchez D, Malek NP, Stange EF, Wehkamp J. Crohn's disease-derived monocytes fail to induce Paneth cell defensins. *Proc Natl Acad Sci U S A*. 2015;112(45):14000–14005.
11 Torres J, Mehandru S, Colombel JF, Peyrin-Biroulet L. Crohn's disease. *Lancet*. 2017;389(10080):1741–1755.
12 Sazonovs A, Stevens CR, Venkataraman GR, et al. Large-scale sequencing identifies multiple genes and rare variants associated with Crohn's disease susceptibility. *Nat Genet*. 2022;54(9):1275–1283.
13 Jostins L, Ripke S, Weersma RK, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*. 2012;491(7422):119–124.
14 Ng SC, Tang W, Leong RW, et al. Environmental risk factors in inflammatory bowel disease: a population-based case-control study in Asia-Pacific. *Gut*. 2015;64(7):1063–1071.
15 VanDussen KL, Liu TC, Li D, et al. Genetic variants synthesize to produce paneth cell phenotypes that define subtypes of Crohn's disease. *Gastroenterology*. 2014;146(1):200–209.
16 Liu TC, Naito T, Liu Z, et al. LRRK2 but not ATG16L1 is associated with Paneth cell defect in Japanese Crohn's disease patients. *JCI Insight*. 2017;2(6):e91917.
17 Liu TC, Kern JT, VanDussen KL, et al. Interaction between smoking and ATG16L1T300A triggers Paneth cell defects in Crohn's disease. *J Clin Invest*. 2018;128(11):5110–5122.
18 Liu TC, Kern JT, Jain U, et al. Western diet induces Paneth cell defects through microbiome alterations and farnesoid X receptor and type I interferon activation. *Cell Host Microbe*. 2021;29(6):988–1001.e1006.
19 Cadwell K, Patel KK, Maloney NS, et al. Virus-plus-susceptibility gene interaction determines Crohn's disease gene Atg16L1 phenotypes in intestine. *Cell*. 2010;141(7):1135–1145.
20 Balasubramanian I, Bandyopadhyay S, Flores J, et al. Infection and inflammation stimulate expansion of a CD74(+) Paneth cell subset to regulate disease progression. *EMBO J*. 2023;42(21):e113975.

21 Khaloian S, Rath E, Hammoudi N, et al. Mitochondrial impairment drives intestinal stem cell transition into dysfunctional Paneth cells predicting Crohn's disease recurrence. *Gut*. 2020;69(11):1939–1951.
22 Alharbi E, Rajaram A, Cote K, et al. A deep learning-based approach to estimate Paneth cell granule area in celiac disease. *Arch Pathol Lab Med*. 2024;148(7):828–835.
23 van der Laak J, Litjens G, Ciompi F. Deep learning in histopathology: the path to the clinic. *Nat Med*. 2021;27(5):775–784.
24 Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng*. 2018;2(10):719–731.
25 Sun L, Marsh JN, Matlock MK, et al. Deep learning quantification of percent steatosis in donor liver biopsy frozen sections. *eBioMedicine*. 2020;60:103029.
26 Marsh JN, Matlock MK, Kudose S, et al. Deep learning global glomerulosclerosis in transplant kidney frozen sections. *IEEE Trans Med Imaging*. 2018;37(12):2718–2728.
27 Marsh JN, Liu TC, Wilson PC, Swamidass SJ, Gaut JP. Development and validation of a deep learning model to quantify glomerulosclerosis in kidney biopsy specimens. *JAMA Netw Open*. 2021;4(1):e2030939.
28 Ohara J, Nemoto T, Maeda Y, Ogata N, Kudo SE, Yamochi T. Deep learning-based automated quantification of goblet cell mucus using histological images as a predictor of clinical relapse of ulcerative colitis with endoscopic remission. *J Gastroenterol*. 2022;57(12):962–970.
29 Liu X, Prasath S, Siddiqui I, et al. Machine learning-based prediction of pediatric ulcerative colitis treatment response using diagnostic histopathology. *Gastroenterology*. 2024;166(5):921–924.e924.
30 Schneider CA, Rasband WS, Eliceiri KW. NIH image to ImageJ: 25 years of image analysis. *Nat Methods*. 2012;9(7):671–675.
31 Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: *Medical image computing and computer-assisted intervention – MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. New York, NY: Springer Berlin Heidelberg: 2015:234–241.
32 van der Walt S, Schonberger JL, Nunez-Iglesias J, et al. scikit-image: image processing in Python. *PeerJ*. 2014;2:e453.
33 Bankhead P, Loughrey MB, Fernandez JA, et al. QuPath: open source software for digital pathology image analysis. *Sci Rep*. 2017;7(1):16878.
34 Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825–2830.
35 Perminow G, Beisner J, Koslowski M, et al. Defective Paneth cell-mediated host defense in pediatric ileal Crohn's disease. *Am J Gastroenterol*. 2010;105(2):452–459.
36 Wehkamp J, Wang G, Kubler I, et al. The Paneth cell alpha-defensin deficiency of ileal Crohn's disease is linked to Wnt/Tcf-4. *J Immunol*. 2007;179(5):3109–3118.
37 Wehkamp J, Salzman NH, Porter E, et al. Reduced Paneth cell alpha-defensins in ileal Crohn's disease. *Proc Natl Acad Sci U S A*. 2005;102(50):18129–18134.
38 Levine JE, Huber E, Hammer ST, et al. Low Paneth cell numbers at onset of gastrointestinal graft-versus-host disease identify patients at high risk for nonrelapse mortality. *Blood*. 2013;122(8):1505–1509.
39 Gunther C, Martini E, Wittkopf N, et al. Caspase-8 regulates TNF-alpha-induced epithelial necroptosis and terminal ileitis. *Nature*. 2011;477(7364):335–339.
40 Gunther C, Ruder B, Stolzer I, et al. Interferon lambda promotes Paneth cell death via STAT1 signaling in mice and is increased in inflamed ileal tissues of patients with Crohn's disease. *Gastroenterology*. 2019;157(5):1310–1322.e13.
41 Ma C, Haritunians T, Gremida AK, et al. Ileal Paneth cell phenotype is a cellular biomarker for pouch complications in ulcerative colitis. *J Crohns Colitis*. 2024;2:jjae105. https://doi.org/10.1093/ecco-jcc/jjae105.