



**QUEEN'S  
UNIVERSITY  
BELFAST**

## **Development of a semi-automated method for tumour budding assessment in colorectal cancer and comparison with manual methods**

Fisher, N. C., Loughrey, M. B., Coleman, H. G., Gelbard, M. D., Bankhead, P., & Dunne, P. D. (2022). Development of a semi-automated method for tumour budding assessment in colorectal cancer and comparison with manual methods. *Histopathology*, 80(3), 485-500. <https://doi.org/10.1111/his.14574>, <https://doi.org/10.1111/his.14574>

**Published in:**  
Histopathology

**Document Version:**  
Peer reviewed version

**Queen's University Belfast - Research Portal:**  
[Link to publication record in Queen's University Belfast Research Portal](#)

### **Publisher rights**

Copyright 2021, John Wiley & Sons Ltd.  
This work is made available online in accordance with the publisher's policies. Please refer to any applicable terms of use of the publisher.

### **General rights**

Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

### **Take down policy**

The Research Portal is Queen's institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact [openaccess@qub.ac.uk](mailto:openaccess@qub.ac.uk).

### **Open Access**

This research has been made openly available by Queen's academics and its Open Research team. We would love to hear how access to this research benefits you. – Share your feedback with us: <http://go.qub.ac.uk/oa-feedback>

MISS NATALIE C FISHER (Orcid ID : 0000-0003-1769-7468)

DR MAURICE LOUGHREY (Orcid ID : 0000-0001-8424-1765)

DR PETER BANKHEAD (Orcid ID : 0000-0003-4851-8813)

Article type : Original Article

## **Development of a semi-automated method for tumour budding assessment in colorectal cancer and comparison with manual methods**

Running title: Semi-automated tumour budding assessment

Natalie C Fisher<sup>1\*</sup>, Maurice B Loughrey<sup>1,2,3\*</sup>, Helen G Coleman<sup>1,3</sup>, Melvin D Gelbard<sup>5</sup>, Peter Bankhead<sup>4,5,6\*</sup> and Philip D Dunne<sup>1\*</sup>.

<sup>1</sup>The Patrick G Johnston Centre for Cancer Research, Queen's University Belfast, UK.

<sup>2</sup>Department of Cellular Pathology, Belfast Health and Social Care Trust, Belfast, UK.

<sup>3</sup>Centre for Public Health, Queen's University Belfast, UK.

<sup>4</sup>Edinburgh Pathology, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK

<sup>5</sup>Centre for Genomic & Experimental Medicine, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK

<sup>6</sup>Cancer Research UK Edinburgh Centre, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK

\*These authors contributed equally to this work

Corresponding author:

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the [Version of Record](#). Please cite this article as [doi: 10.1111/HIS.14574](https://doi.org/10.1111/HIS.14574)

This article is protected by copyright. All rights reserved

Accepted Article

Dr. Maurice Loughrey  
Department of Cellular Pathology,  
Institute of Pathology,  
Royal Victoria Hospital,  
Grosvenor Road  
Belfast BT12 6BA  
N. Ireland  
Tel. +442890 633224

Disclosure: The authors have no conflicts of interest to declare

Word count: 4825

## Abstract

**Aims:** Tumour budding is an established prognostic feature in multiple cancers but is not routinely assessed in pathology practice. Efforts to standardise and automate assessment have shifted from haematoxylin and eosin (H&E)-stained images towards cytokeratin immunohistochemistry. In this study, we compare manual H&E and cytokeratin assessment methods with a semi-automated approach built within QuPath open-source software.

**Methods and Results:** Budding was assessed in cores from the advancing tumour edge in a cohort of stage II/III colon cancers (n=186). The total numbers of buds detected by each method were: manual H&E (n=503), manual cytokeratin (n=2290), semi-automated (n=5138). More than four times the number of buds were detected manually using cytokeratin compared to H&E. 1734 individual buds were identified on both manual and semi-automated assessments applied to cytokeratin images, representing 75.7% of the buds identified manually (n=2290) and 33.7% of the buds detected using the semi-automated method (n=5138). Higher semi-automated bud scores were due to any discrete area of cytokeratin immunopositivity within an accepted area range being identified as a bud, regardless of shape or crispness of definition, and to inclusion of tumour cell clusters within glandular lumina ("luminal pseudobuds"). Although absolute numbers differed, semi-automated and manual bud counts were strongly correlated across cores ( $p=0.81$ ,  $p<0.0001$ ). All methods of budding assessment demonstrated poorer survival associated with higher budding scores.

**Conclusions:** We present a new QuPath-based approach to tumour budding assessment, which compares favorably to established methods and offers a freely-available, rapid and transparent tool that is also applicable to whole slide images.

**Keywords:** tumour budding; colorectal cancer; cytokeratin immunohistochemistry; machine learning; QuPath.

## Introduction

Tumour budding (TB) is the histological manifestation of local tumour cell dissemination, usually most evident at the invasive front region of a tumour mass. TB is an established prognostic factor in a number of solid tumours [1], although it has been most extensively studied in colorectal cancer (CRC). In pT1 CRC, the presence and extent of TB is predictive for nodal metastatic disease, and thus can be used as a clinical tool for identifying patients most likely to benefit from surgical resection [2]. TB has also been shown to have prognostic value in all other stages of CRC, with most evidence reported for stage II disease [1,3,4].

Despite the potential clinical utility of TB, inconsistent qualitative criteria, definitions and non-standardised reporting have proven an obstacle to routine implementation in pathology practice and TB generally remains a “non-core” item in CRC reporting datasets [5–7]. In an attempt to address this issue in 2016, the International Tumour Budding Consensus Conference (ITBCC) established a consensus definition of a tumour bud, namely a single tumour cell or tumour cell cluster of up to four cells, and an agreed histopathological method of assessment [8]. Although encouraging data was emerging at that time regarding TB assessment by cytokeratin (CK) immunohistochemistry (IHC), most of the established evidence was based on haematoxylin and eosin (H&E) assessment. The consensus preference from ITBCC was for H&E staining in conjunction with a three-tier scoring system within a “hot spot” field area normalised to 0.785 mm<sup>2</sup>.

Since emergence of the consensus budding definition from ITBCC, there has been increased focus on standardisation, reproducibility and automation, with a view to clinical implementation. This was the subject of a recent comprehensive review, which summarised twelve publications describing differing semi-automated approaches to TB assessment, almost all applied to CRC [9]. Most used commercially-available software but two utilised open-source software (ImageJ), and some used a form of machine

learning. Importantly, almost all were applied to CK IHC images, with only one method proposed for H&E. Other groups pursuing manual rather than semi-automated assessment of TB have also advocated for a CK IHC-based approach [10]. However, a recent expert Delphi consensus process addressing TB concluded that more evidence was required before incorporating IHC into TB scoring [11].

One advantage of CK IHC over H&E assessment is the potential for greater reproducibility in overall TB grade [12], addressing a limiting step in progressing TB towards clinical implementation. While most studies have compared only overall TB grade, very few studies have examined TB assessment at the individual bud level, which is likely where most discordance lies. Recently, Bokhorst *et al* compared evaluation by a panel of seven ITBCC experts of 3000 candidate buds from CK-stained sections representing 46 patients with CRC and found only moderate agreement [13]. Consensus classification was not reached on 41% of the candidate buds. Agreement was slightly better in this study for H&E assessment of individual buds compared with CK IHC, but far fewer H&E candidate buds were presented for evaluation.

In the current study, we compare manual H&E and CK assessment methods with a new semi-automated approach to TB assessment performed on digital images from a cohort of stage II and III colon cancers. Manual and semi-automated annotation of individual candidate buds on the same CK IHC images allowed scrutiny of discordance at the individual bud level and consideration of the optimal definition of a tumour bud for these methods of assessment. Results were analysed for all methods against impact on survival, as a measure of relative performance and comparison of potential clinical utility.

## Materials and methods

### STUDY COHORT

The study utilised an established Northern Ireland population-based resource of 661 stage II and III colon cancers, creation of which has been fully described previously (Northern Ireland Biobank ethical approval references NIB13-0069/87/88 and NIB20-0334) [14]. The resource includes tissue microarrays (TMA), generated from representative tumour blocks containing the tumour advancing edge, with one 1 mm diameter core per tumour taken from a random area along the advancing edge. Although this does not reflect clinical practice, where TB grade is based on the “hotspot” area from within a representative whole tumour section, use of TMAs in this study allowed high throughput and representation of the full morphological spectrum of colon cancer.

3  $\mu\text{m}$  sections from each TMA were stained with H&E and with an anti-cytokeratin immunohistochemical antibody (Cam5.2; Ventana, mouse monoclonal, Cell Conditioning 1 for 8 min, DAB chromogen) on a BenchMark ULTRA (Ventana Medical Systems Inc.) automated slide stainer. Glass slides were scanned on an Aperio AT2 Scanner (Leica Biosystems, Newcastle, United Kingdom) at x40 and imported into the open-source software QuPath (v0.2.3) [15] for evaluation. The scanned TMA images are available from the Northern Ireland Biobank [16] upon application.

The suitability of individual CK IHC-stained cores for inclusion was determined by manual visual assessment of the scanned images, after application of the QuPath TMA dearraying tool. Of note, TMA sampling from the advancing tumour edge is likely to generate a significant number of “misses”, with only peritumoural tissue sampled. Of the 486 cores with sufficient tumour present and matched clinicopathological data, individual cores were also excluded if (a) only mucinous or signet ring cell carcinoma was present (n=26), (b) there were large areas of tumour necrosis (n=26) (c) tumour present exhibited weak, patchy or negative immunostaining (n=82), (d) there was significant stromal CK

immunopositivity (n=25), or (e) tissue folding, fragmentation or any other technical artefacts precluded assessment (n=72) (Supplementary Figure 1). After the above exclusions, 255 cores remained for CK IHC evaluation. Manual H&E assessment for inclusion was performed after CK IHC assessment, and a further 61 cores were excluded, due to either a lack of tumour or tissue artefacts as described above, precluding H&E assessment. A further eight cases with less than one month of follow-up time were also excluded from the analysis. This left 186 cases for analysis, having comparative TB data for all four methods of assessment, as detailed below, and clinicopathological data available including sufficient follow-up.

#### MANUAL BUDDING ASSESSMENT

Buds were manually assessed on H&E and CK IHC images by an expert gastrointestinal pathologist (MBL). This process is depicted in Figure 1A-1E. Within QuPath, after dearranging, individual cores were shrunk by 30  $\mu\text{m}$  to correlate with semi-automated assessment in excluding candidate buds touching the periphery of the core. Each individual bud was manually marked on all images using the point tool within QuPath, enabling quick and accurate quantification per core and the ability to review each individual bud counted. The ITBCC recommendations for H&E TB assessment were followed, with the only exception being that the TMA cores did not represent the budding “hot spots” for each tumour. However, each 1 mm diameter core approximates the ITBCC recommended 0.785 mm<sup>2</sup> area for TB assessment [8]. Furthermore, by using random cores from the advancing edge our analyses were tested in a wide range of morphological conditions. Pre-determination of the tumour region for assessment with the TMA approach allowed inter-method comparison of individual buds. “Pseudobuds” within areas of heavy acute inflammation were excluded as recommended [8,11].

For initial manual assessment of CK-stained cores, the aim was to annotate as buds clusters of up to four tumour cells, as on H&E, accepting that visualising and counting tumour cell nuclei is more difficult on CK IHC than on H&E (Figure 1C). Regions of irregular or ill-defined IHC staining were excluded, some considered likely to represent



cellular fragments rather than viable buds. After this initial assessment was complete, annotated buds (CK all) were reassessed by the same observer to apply the recently suggested additional criterion of nuclear pallor in defining a bud [13]. Those single cells or clusters lacking an identifiable region of nuclear pallor were removed to generate an additional budding dataset (CK pallor) which excluded objects lacking this potentially important feature (Figure 1E).

### SEMI-AUTOMATED BUDDING ASSESSMENT

The semi-automated method was based on a binary (immunopositive/immunonegative) threshold classifier built within QuPath (v0.2.3) and applied to the CK IHC images to identify tumour epithelium. This process is depicted in Figure 1F-1J. As before, following dearraying, individual cores were shrunk by 30  $\mu\text{m}$  to exclude candidate buds touching the periphery of the core. All lumens completely encapsulated by positive staining were filled in, to prevent the detection of luminal tumour cells or cellular fragments mimicking buds (“luminal pseudobuds”) (Figure 1G). A pixel classifier was created in QuPath to identify connective discrete areas of immunopositivity by combining image downsampling, stain separation using colour deconvolution [17], Gaussian smoothing and global thresholding within a single step (Resolution: 1.86 $\mu\text{m}/\text{px}$ ; Channel: DAB; Prefilter: Gaussian; Smoothing sigma: 1.0; Threshold: 0.4). Buds were defined using this method not by number of tumour nuclei, but by area of CK immunopositivity. An acceptable range of bud area was derived from analysis of the range of areas of the manually annotated CK buds (described in detail below). Those objects with areas outside this range were excluded as buds (Figure 1H-1J).

### STATISTICAL ANALYSIS

Cox PH was conducted in Stata version 16 (Timberlake Consultants, StataCorp, College Station, TX, USA). All other analysis was conducted in R 4.0.3 (R Foundation for

Statistical Computing, Vienna, Austria) [18]. Statistical differences between the clinicopathological characteristics of the subset of patients utilised in this study compared to the overall cohort were determined. The Wilcoxon rank-sum test was applied to those groups with two levels, and Pearson's Chi-squared test without continuity correction or Fisher's exact test was applied to categorical variables where appropriate. The Kruskal-Wallis rank-sum test was used for the continuous variable.

Descriptive statistics were performed on the number of tumour buds detected per tissue core by each of the scoring methods. Spearman's correlation coefficient was used to determine the strength of the linear relationship between each of the scoring methods.

Univariable and multivariable analyses using the Cox proportional hazards regression model were performed to calculate hazard ratios (HR) and 95% confidence intervals (CI) for overall survival according to TB. Adjusted models were tested for family history of CRC, tumour grade (differentiation) and MSI status but these factors were excluded since they did not influence the model. Multivariable adjustments were age (<50, 50-<60, 60-<70, 70-<80, ≥80 years), sex (male, female), adjuvant chemotherapy receipt (yes, no), stage (II,III) and ECOG performance status (0-1, 2, 3-4, unknown). As the TMA cores in this study represent random cores from the tumour advancing edge, rather than TB hotspots, the ITBCC three category cut-offs are not strictly applicable. Therefore, survival analysis was conducted in two ways: (i) based on continuous bud counts to maximise statistical power, with per increment increases for each method based on relative ratios of total bud counts between methods; and (ii) applying modified ITBCC cut-offs to mimic categorisation of scores for clinical decision making, and to generate Kaplan-Meier curves of prognostication. ITBCC three category cut-offs were utilised for H&E scores (≤4, 5-9, ≥10 buds) and cut-offs for the other methods scaled up according to the TB score distribution for each method.

## Results

Of the original cohort, 186 individual cases were included in the study analysis. The overall clinicopathological characteristics are summarised in Table 1, which demonstrates that the subset of patient samples used in this current study shows no meaningful differences when compared to the overall stage II/III population-based cohort and can be considered a representative subset for analysis.

### *Deriving bud area range for semi-automated method*

Semi-automated bud counts first required definition of an acceptable range of bud area, derived from analysis of the range of areas of the manually annotated CK buds. The semi-automated method initially identified all discrete areas of CK immunopositivity. Immunopositive areas, representing candidate buds, were initially captured over a wide size range (5-3000  $\mu\text{m}^2$ ). Extremely small areas represented either tiny immunopositive tumour fragments, often in the context of gland rupture, (Figure 2A&2B) or non-specific immunostaining of uncertain nature (Figure 2C&2D). Large tumour areas were also annotated. By mapping the manual CK annotations to the semi-automated annotations, the areas of all manually annotated CK buds (CK all) could be measured within QuPath (Figure 2E&2F) and exported for analysis. The median CK bud area of the manually annotated CK buds (CK all), as measured by QuPath, was 225  $\mu\text{m}^2$  (Figure 3A; interquartile range 133-388  $\mu\text{m}^2$ ). The images, including manual and semi-automated annotations, of outliers at the low and high end of the area scale were reviewed, to explain implausibly small and large areas for some manually annotated buds. In some single cell buds, the semi-automated method excluded from the area measurement a prominent region of central nuclear pallor, thereby underestimating the true bud area (Figure 2G&2H). For some closely approximated buds, QuPath failed to resolve these as separate buds and considered their total combined area as a single immunopositive region, resulting in an apparent manually detected bud with a large area (Figure 2I&2J). Taking these erroneous extreme values into consideration, a range of 40-700  $\mu\text{m}^2$  was chosen as acceptable in this study for defining a bud based on area of CK

immunopositivity. Applying this definition, Figure 3B demonstrates by histogram the resultant areas and frequencies of the buds detected by the semi-automated method, having a lower modal bud area compared to the manual CK (all) method.

#### *Total bud count comparisons*

The total number of buds detected by each method (Figure 4A), over the 186 TMA cores, were as follows; manual H&E (n=503), CK all (n=2290), CK pallor (n=1825) and semi-automated (n=5138). These findings indicate that more than four times the number buds were detected using CK (CK all) compared to H&E, and more than three times the number if restricting to those buds with central pallor (CK pallor). The semi-automated method detected over ten times more buds than H&E and over twice as many buds as CK (CK all). Comparing bud totals and frequencies for each method showed progressively increasing numbers of cases with higher numbers of buds moving from H&E to CK to semi-automated assessments (Figure 4B). Comparison of total bud numbers between H&E and CK showed moderate correlation (Figure 4C,  $\rho=0.60$ ,  $p<0.0001$ ), whereas strong correlation was observed between CK all and semi-automated methods (Figure 4D,  $\rho=0.81$ ,  $p<0.0001$ ).

#### *Bud by bud comparisons*

As both manual CK assessments and the semi-automated assessment were performed on the same set of images, bud by bud comparison was possible for these methods. A total of 1734 individual buds were identified both by manual assessment (CK all) and semi-automated detection, representing 75.7% of the total manual buds identified (n=2290) and 33.7% of the total semi-automated buds detected (n=5138) (Figure 5). Accepting the manual CK method as the relevant gold standard, these equate to the sensitivity and positive predictive value respectively of the semi-automated method for detection of CK (CK all) buds.

#### *Bud discordance between methods*

Many tumour areas demonstrated excellent concordance, with buds being detected by both manual CK and semi-automated assessment methods after application of the specified area range for the semi-automated method (Figure 6A&6B). However, elsewhere concordance between these assessment methods was poor. This was in large part due to the semi-automated method accepting as a bud any discrete area of CK immunopositivity within the accepted area range, regardless of shape or crispness of definition, features which would typically be considered in the manual assessment of a bud (Figure 6C&6D). The other main explanation for much greater numbers of buds by the semi-automated method relates to “luminal pseudobuds”. Manual assessment discounts as buds, tumour cells or clusters lying within glandular lumina. When surrounded by circumferential staining, QuPath was able to fill in the glandular lumina, to avoid counting such mimics as buds (Figures 1F&1G, 6E&6F). However, when staining was not circumferential, QuPath counted these luminal immunopositive fragments as buds (Figure 6G&6H). This was a particular problem at core peripheries, where the complete gland circumference was not captured within the core (Figure 6I&6J). The inclusion of the more stringent nuclear pallor criterion to define a CK bud by manual assessment had a minor additional impact on the discordance in bud numbers between manual CK and semi-automated assessments (Figure 5).

A smaller number of manual CK buds (CK all and CK pallor) were not detected by the semi-automated method. These are explained by erroneous bud area measurement, as described above. Incorrect assessment of true bud area, because of exclusion of a region of nuclear pallor (Figure 2G&2H) or failure to resolve closely adjacent buds (Figure 2I&2J), generated areas below or above the accepted range, and thereby failure to identify these manually detected buds by the semi-automated method.

### *Survival analysis*

Of the 186 patients included in the analysis, by the end of follow-up (mean  $\pm$  standard deviation,  $5.5 \pm 3.0$  years; range 0.12-10 years, interquartile range 2.89-8.19 years), 90 had died of which 60 were from a CRC-related cause. All four methods of TB assessment demonstrated reduced survival associated with higher budding scores (Table 2). HRs

were similar for both of the CK methods and for the semi-automated method in the univariable (manual CK all: HR 1.09, 95%CI 1.05-1.14; manual CK pallor: HR 1.11, 95%CI 1.06-1.18; semi-automated: HR 1.09, 95%CI 1.04-1.14) and multivariable (manual CK all: HR 1.06, 95%CI 1.02-1.11; manual CK pallor: HR 1.08, 95%CI 1.02-1.14; semi-automated: HR 1.06, 95%CI 1.01-1.11) models, and slightly lower for the H&E method in both univariable (HR 1.03, 95%CI 1.01-1.05) and multivariable (HR 1.02, 95%CI 1.00-1.04) models. All findings were statistically significant aside from H&E findings in the multivariable model.

Kaplan-Meier survival analysis showed patients with higher TB grades had reduced overall five year survival, when assessed by any of the four methods presented (Figure 7). Stratification was not significant for H&E assessment ( $p=0.14$ ) but was significant for the other three methods, all of which showed comparable stratification ( $p=0.00016$ ,  $p=0.00014$ ,  $p=0.0011$ ). Introduction of nuclear pallor to the manual CK assessment did not meaningfully impact stratification.

## Discussion

TB is well established as an adverse prognostic feature in CRC in several clinical settings [1]. Despite considerable existing evidence in this regard, assessment of TB has not yet been incorporated into routine clinical practice. In large part, this is because of uncertainty regarding the most appropriate method of assessment, specifically the most appropriate stain for counting buds and whether to persist with manual assessment or adopt some form of semi-automated approach. In this study, we used QuPath to develop a new digital pathology-based semi-automated TB assessment tool for CK-stained sections, which we then compared to established methods of TB assessment in a cohort of colon cancers using a TMA approach. As the study included TMA cores from the tumour advancing edge of stage II/III colon cancers, rather than the budding hotspot advocated for clinical use, the primary focus of this paper was a bud by bud comparison of manual CK and our semi-automated assessment method, rather than to provide further evidence of adverse prognostic significance of TB.

Our data indicates that CK IHC detected over four times more buds than H&E-based assessment of parallel sections, which is consistent with previous studies observing three to six times more buds with CK IHC than with H&E staining [12]. Although not examined in this study, it is postulated that CK IHC is particularly valuable in highlighting single cell buds and distinguishing these from epithelioid stromal or histiocytic cells by indicating their epithelial cell lineage, less readily apparent on H&E. Bokhorst *et al* have hypothesised that inter-observer variability on H&E assessment may be more problematic for single cell buds than for two to four cell buds [13]. H&E assessment allows better evaluation of the microenvironment surrounding buds and so it is possible that a further reason contributing to fewer H&E buds relates to greater exclusion of so-called pseudobuds at sites of active inflammation, often related to gland rupture [1]. The inflammatory environment is less readily appreciated in CK IHC preparations, meaning pseudobuds may be less identifiable and therefore less likely to be excluded.

The threshold semi-automated approach identified approximately 2.5 times more buds than manual CK assessment. Higher bud counts have been observed previously when comparing a semi-automated to manual CK assessment method, but without quantification [19]. In data presented here, we find that bud by bud comparison revealed only moderate agreement between these two assessment methods for individual buds. Some of the discrepancy might be explained by the tendency of any human observer to err slightly on the side of under-counting, either through occasionally missing a possible true bud or by making a conservative judgement in an ambiguous case. By contrast, one can expect a threshold-based approach, calibrated to identify true buds based upon CK immunopositivity, to err definitively on the side of overestimation, because it will consistently include more irregular or ill-defined ambiguous tumour cell clusters of a defined size. It is possible that incorporating further criteria into the bud definition may improve agreement between semi-automated and manual assessments, such as a measure of circularity [20]. However, given that there is no *a priori* reason to suppose buds are circular, this can introduce further subjectivity. In this study we have aimed to minimise the adjustable parameters, relying primarily upon a staining threshold and area filter to achieve a replicable baseline of quantitative assessment. The area range we selected to define a tumour bud (40-700  $\mu\text{m}^2$ ) was based on the corresponding area range of manually detected CK buds, which is wider than that chosen by Takamatsu *et al.* (100-480 $\mu\text{m}^2$ ) but narrower than that chosen by Bokhorst *et al.* (25-1000 $\mu\text{m}^2$ ) [13,20]. This already indicates the lack of accepted parameters in defining bud characteristics through image analysis, although such parameters will inevitably have a profound influence upon the absolute numbers of buds detected. Interestingly, we found that, despite the substantial differences in absolute bud counts between methods of assessment, correlation remained high – suggesting that the signal remains high amidst the noise.

As there is evidence to support high TB as an adverse prognostic factor across all stages of CRC [1,3,4], survival analysis was conducted applying the four methods of TB assessment, as a measure of comparative performance. Despite the limitations of random core sampling, TB assessed by all four methods was, as expected, significantly



associated with reduced overall survival at five years of follow-up. This association was weakest for H&E assessment, and non-significant on the multivariable model, but it is likely that H&E assessment, with the lowest bud counts in general, will have been impacted more by the random core approach in our study in comparison to the other methods yielding much higher bud counts. Nevertheless, the other three methods all stratified patients better than H&E with respect to survival and achieved almost identical hazard ratios based on evaluation of continuous bud counts. Importantly, despite its simplicity and only moderate agreement with manual cytokeratin assessment for individual buds, the semi-automated threshold approach in QuPath provided an association between higher grades of TB and worse overall patient survival, even when applied to random tumour cores.

A recent modified Delphi process conducted amongst an international group of expert gastrointestinal pathologists supported ongoing assessment of TB using H&E-stained slides, with more evidence required to move to IHC, but also suggested that digital image analysis was likely to facilitate implementation into clinical practice [11]. As almost all TB algorithms published to date rely on CK rather than H&E-stained images, it seems likely that the optimal approach will ultimately be one based on evaluation of the most representative tumour section, stained for CK. With increasing developments in digital pathology and growing access to digital whole slide images in routine practice, some form of semi-automated approach is attractive for reasons of efficiency, cost and reproducibility. Such semi-automated methods can be easily applied over a much larger tumour area to accurately identify the budding density over any agreed area denominator. The consensus 0.785 mm<sup>2</sup> area applicable to microscopy is less relevant to whole slide image analysis. Nevertheless, most current evidence for TB significance is based on this hotspot area, and correlation with microscopy assessment of TB will be important for the foreseeable future.

It is likely that the semi-automated approach to budding assessment described in this study is overly simplistic for clinical use as it is unable to detect some of the more subtle

morphological features of tumour buds, such as nuclear pallor, nor exclude mimics such as pseudobuds. Future clinical implementation will require more refined methodologies, likely involving deep learning [9,21], however as yet no such method is widely available to the TB community. The semi-automated QuPath approaches developed and applied in this study will be of potential benefit to ongoing translational TB research in retrospective cohorts as a much cheaper, more efficient and readily customisable open-source method compared to commercial software solutions. Such tools can be utilised either as a standalone TB assessment or as an adjunct to developing more sophisticated methods for example by identifying large numbers of candidate buds for consensus expert evaluation, classification and application to training of deep learning algorithms.

Assessment of TB by CK IHC has been shown by some studies to improve inter-observer reproducibility, an important requirement when considering incorporation of any new parameter into routine pathology practice [12,22]. However, a recent study employing CK IHC for TB assessment examined inter-observer agreement at the individual bud level and found only moderate agreement, no better than for H&E assessment [13]. The authors considered two reasons for this: firstly, that individual tumour nuclei within immunopositive clusters are sometimes difficult to discern, and therefore count, on CK IHC; and secondly, that the surrounding inflammatory environment is more difficult to assess on CK IHC than on H&E, making evaluation of potential “pseudobudding” more challenging. Less evidence is available on reproducibility of semi-automated methods but it is intuitive that more automation implies greater reproducibility. Takamatsu *et al.* found significantly better reproducibility amongst three pathologists with their semi-automated method (kappa coefficient = 0.781) compared to manual assessment (kappa coefficient = 0.463) [20]. Nevertheless, some degree of manual oversight remains important whilst new methods are developed and tested.

Introducing the additional criterion of nuclear pallor into the manual CK assessment method made no meaningful alteration to the resultant hazard ratio (CK pallor HR 1.11, 95% CI 1.06-1.18; CK all HR 1.09, 95% CI 1.05-1.14) or the Kaplan-Meier survival

stratification, providing no real evidence from this study for inclusion of this criterion. Previously suggested by Bokhorst *et al.* [13], to help exclude CK positive non-viable tumour cell fragments from consideration as buds, this feature should be the focus of future studies based on hotspot TB assessment on whole tumour sections from appropriate CRC cohorts, to ascertain the potential impact of this morphological criterion on clinical relevance of TB and inform future discussions on bud definition.

This study is limited by the random nature of the tumour core samples, limiting analysis of the clinical significance of TB scores with respect to survival analyses, and by the single pathologist manual assessments without any ability to assess reproducibility. However, a detailed comparison of different TB assessment methods is described, applied to a wide morphological spectrum of colon cancers, with bud by bud comparison between methods.

Although our CK thresholding approach resembles methods applied in previous TB studies [9,20,23], to our knowledge the current study is the first to describe an interactive tool for TB assessment that is freely available, open-source, and can be readily applied to whole slide images as part of a full analysis workflow. This is possible because of the extensive additional functionality within QuPath, including the ability to precisely define regions of interest (e.g. a 1 mm boundary delineating the tumour advancing edge), identify hotspots, and export quantitative metrics. These features are illustrated in Figure 8, applying the methods adopted in this study to a whole slide image from a sample CRC case rich in tumour buds. Manually-derived and semi-automated budding density “heat maps” are almost identical. In contrast to assessment approaches driven entirely by machine learning, which can be confounded by even subtle variations in staining or scanning [24,25], our comparatively simple thresholding method can be readily adapted to new images by adjusting a small number of intuitive parameters – making it immediately accessible to any laboratory wishing to apply the technique. Nevertheless, it is clearly desirable to achieve a better discrimination of true buds from false positives. In this regard, QuPath’s generic support for machine learning, previously described for cell

classification [15], can be incorporated into a more elaborate analysis workflow. Having established in this study the first open and replicable end-to-end analysis protocol for TB assessment suitable for whole slide images, we aim to collaborate with other groups to develop a refined, open-source bud identification algorithm based upon a more diverse training dataset across multiple centres.

In conclusion, we present a new, semi-automated, QuPath-based approach to TB assessment. This demonstrates moderate agreement with manual CK-based assessment at a bud-by-bud level and comparable ability to stratify a cohort of patients with stage II/III colon cancer for overall survival. More importantly, it shows QuPath's potential as a freely-available, rapid and transparent tool for TB assessment, applicable to whole slide images, which can be used in translational research as a standalone method or as an aid in developing future approaches suitable for clinical implementation.

## **Acknowledgements**

The samples used in this research were received from the Northern Ireland Biobank which has received funds from HSC Research and Development Division of the Public Health Agency in Northern Ireland and the Friends of the Cancer Centre.

The Belfast Health and Social Care Trust Department of Cellular Pathology is acknowledged for assisting with immunohistochemistry.

## **Author Contributions**

Study concept and design: N.F., M.L., P.B. and P.D.; development of methodology N.F., M.L., P.B. and P.D.; QuPath software development: M.G. and P.B.; acquisition, analysis and interpretation of data and statistical analysis: N.F., M.L., H.C., and P.D; writing initial draft: N.F. and M.L.; review and revision of subsequent drafts, N.F., M.L., H.C., M.G., P.B. and P.D. All authors read and approved the final manuscript.

## **Funding**

The study cohort creation was enabled by funding from Cancer Research UK (ref. C37703/A15333 and C50104/A17592) and a Northern Ireland HSC R&D Doctoral Research Fellowship (ref. EAT/4905/13). This work was supported by the Queens University Belfast Foundation (P.D., N.F.; Musgrave scholarship), a Cancer Research UK early detection grant (P.D.; A29834), a Cancer Research UK Career Establishment Award (H.C.; C37703/A25820) and the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation (awarded to P.B., M.G. was funded by grant number 2019-207148).

## **Data Availability Statement**

QuPath scripts used in this current study are publicly available at <https://github.com/petebankhead/qupath-budding-scripts>. The images and datasets used in the current study are available on application to the Northern Ireland Biobank.

## References

- 1 Lugli A, Zlobec I, Berger MD, et al. Tumour budding in solid cancers. *Nat. Rev. Clin. Oncol.* 2021; **18**; 101-115.
- 2 Bosch S, Teerenstra S, De Wilt JW, et al. Predicting lymph node metastasis in pT1 colorectal cancer: A systematic review of risk factors providing rationale for therapy decisions. *Endoscopy* 2013; **45**; 827-834.
- 3 Landau MA, Zhu B, Akwuole FN, et al. Histopathological Predictors of Recurrence in Stage III Colon Cancer: Reappraisal of Tumor Deposits and Tumor Budding Using AJCC8 Criteria. *Int. J. Surg. Pathol.* 2019; **27**; 147-158.
- 4 Van Wyk HC, Park JH, Edwards J, et al. The relationship between tumour budding, the tumour microenvironment and survival in patients with primary operable colorectal cancer. *Br. J. Cancer* 2016; **115**; 156-163.
- 5 Brown I, Kumarasinghe, Priyanthi, Rosty C, Ellis D, et al. Colorectal Cancer Structured Reporting Protocol. 4th ed. The Royal College of Pathologists of Australasia; 2020.
- 6 Loughrey M, Quirke P, Shepherd NA. Dataset for Histopathological Reporting of Colorectal Cancer. 4th ed. Royal College of Pathologists; 2018.
- 7 Burgart LJ, Kakar S, Shi C, et al. Protocol for the Examination of Resection Specimens From Patients With Primary Carcinoma of the Colon and Rectum. 4.1.0.0. College of American Pathologists; 2020.
- 8 Lugli A, Kirsch R, Ajioka Y, et al. Recommendations for reporting tumor budding in colorectal cancer based on the International Tumor Budding Consensus Conference (ITBCC) 2016. *Mod. Pathol.* 2017; **30**; 1299-1311.
- 9 Studer L, Blank A, Bokhorst J-M, et al. Taking tumour budding to the next frontier - a post International Tumour Budding Consensus Conference (ITBCC) 2016 review. *Histopathology* 2021; **78**; 476-484.
- 10 Koelzer VH, Assarzadegan N, Dawson H, et al. Cytokeratin-based assessment of tumour budding in colorectal cancer: analysis in stage II patients and prospective

diagnostic experience. *J. Pathol. Clin. Res.* 2017; **3**; 171-178.

11 Haddad TS, Lugli A, Aherne S, et al. Improving tumor budding reporting in colorectal cancer: a Delphi consensus study. *Virchows Arch.* March 2021.

12 Koelzer VH, Zlobec I, Berger MD, et al. Tumor budding in colorectal cancer revisited: results of a multicenter interobserver study. *Virchows Arch.* 2015; **466**; 485-493.

13 Bokhorst JM, Blank A, Lugli A, et al. Assessment of individual tumor buds using keratin immunohistochemistry: moderate interobserver agreement suggests a role for machine learning. *Mod. Pathol.* 2020; **33**; 825-833.

14 Gray RT, Cantwell MM, Coleman HG, et al. Evaluation of PTGS2 expression, PIK3CA mutation, aspirin use and colon cancer survival in a population-based cohort study. *Clin. Transl. Gastroenterol.* 2017; **8**; e91-10.

15 Bankhead P, Loughrey MB, Fernández JA, et al. QuPath: Open source software for digital pathology image analysis. *Sci. Rep.* 2017; **7**; 16878.

16 Lewis C, McQuaid S, Clark P, et al. The Northern Ireland biobank: A cancer focused repository of science. *Open J. Bioresour.* 2018; **5**; 1-6.

17 Ruifrok AC, Johnston DA. Quantification of histochemical staining by color deconvolution. *Anal. Quant. Cytol. Histol.* 2001; **23**; 291-299.

18 R Core Team. R: A Language and Environment for Statistical Computing. 2020. Available from: <https://www.R-project.org/>.

19 Jepsen RK, Klarskov LL, Lippert MF, et al. Digital image analysis of pan-cytokeratin stained tumor slides for evaluation of tumor budding in pT1/pT2 colorectal cancer: Results of a feasibility study. *Pathol. Res. Pract.* 2018; **214**; 1273-1281.

20 Takamatsu M, Kawachi H, Yamamoto N, et al. Immunohistochemical evaluation of tumor budding for stratifying T1 colorectal cancer: optimal cut-off value and a novel computer-assisted semiautomatic method. *Mod. Pathol.* 2019; **32**; 675-683.

21 Pai RK, Hartman D, Schaeffer DF, et al. Development and initial validation of a deep learning algorithm to quantify histologic features in colorectal carcinoma

including tumour budding/poorly differentiated clusters. *Histopathology* 2021.

22 Kai K, Aishima S, Aoki S, et al. Cytokeratin immunohistochemistry improves interobserver variability between unskilled pathologists in the evaluation of tumor budding in T1 colorectal cancer. *Pathol. Int.* 2016; **66**; 75-82.

23 Vranes V, Rajković N, Li X, et al. Size and shape filtering of malignant cell clusters within breast tumors identifies scattered individual epithelial cells as the most valuable histomorphological clue in the prognosis of distant metastasis risk. *Cancers (Basel)*. 2019; **11**; 1-16.

24 Cui M, Zhang DY. Artificial intelligence and computational pathology. *Lab. Investig.* 2021; **101**; 412-422.

25 Janowczyk A, Zuo R, Gilmore H, et al. HistoQC: An Open-Source Quality Control Tool for Digital Pathology Slides. *JCO Clin. Cancer Informatics* 2019; 1-7.



**Table 1.** Clinicopathological characteristics for the entire patient cohort and for the subset for study analysis.

Characteristic (n=entire/subset)	Entire cohort (n (%))	Subset for analysis (n (%))	p-value
<b>Sex (n=661/186)</b>			0.26
Male	358 (54.2)	92 (49.5)	
Female	303 (45.8)	94 (50.5)	
<b>Age, years (n=661/186)</b>	73 (IR 64-79)	74 (IR 65-80)	0.18
<b>ECOG performance status (n=410/113)</b>			0.86
0-1	338 (82.4)	92 (81.4)	
2	42 (10.2)	11 (9.7)	
3-4	30 (7.3)	10 (8.8)	
<b>Tumour stage (n=661/186)</b>			0.37
II	394 (59.6)	104 (55.9)	
III	267 (40.4)	82 (44.1)	
<b>*Tumour location (n=661/186)</b>			0.42
Proximal	375 (56.7)	116 (62.4)	
Distal	280 (42.4)	69 (37.1)	
Colon unspecified	6 (0.9)	1 (0.5)	
<b>Tumor differentiation (n=657/183)</b>			0.33
Poor	90 (13.7)	20 (10.9)	
Well/moderate	567 (86.3)	163 (89.1)	
<b>Extramural venous invasion (n=610/163)</b>			0.99
Yes	165 (27.0)	44 (27.0)	
No	445 (73.0)	119 (73.0)	
<b>Microsatellite instability status (n=593/175)</b>			0.30
MSI-High	136 (22.9)	47 (26.9)	
MSI-Low	10 (1.7)	5 (2.79)	
Microsatellite stable	447 (75.4)	123 (70.3)	
<b>Adjuvant chemotherapy (n=661/186)</b>			0.85
Yes	186 (28.1)	51 (27.4)	
No	475 (71.9)	135 (72.6)	
<b>Overall survival (n=661/186)</b>			0.64
Alive	354 (53.6)	96 (51.6)	
Dead	307 (46.4)	90 (48.4)	
<b>CRC specific survival (n=566/156)</b>			0.82

Alive	354 (62.5)	96 (61.5)
Dead	212 (37.5)	60 (38.5)

---

\*Distal colonic location defined as splenic flexure and beyond. Rectal cancers were excluded from the study.

ECOG, Eastern Cooperative Oncology Group; MSI, microsatellite instability; CRC, colorectal cancer; IR, interquartile range

**Table 2.** Univariable and multivariable Cox proportional hazards overall survival analysis comparing four methods of tumour budding assessment.

Assessment method	Overall survival	
	Hazard ratio (95% CI)	
	Unadjusted	Multivariable adjusted model
<b>Manual H&amp;E</b>		
Per 1 bud increase	1.03 (1.01-1.05)	1.02 (1.00-1.04)
<i>P-value</i>	<i>0.001</i>	<i>0.08</i>
<b>Manual CK all</b>		
Per 5 bud increase	1.09 (1.05-1.14)	1.06 (1.02-1.11)
<i>P-value</i>	<i>&lt;0.001</i>	<i>0.004</i>
<b>Manual CK pallor</b>		
Per 5 bud increase	1.11 (1.06-1.18)	1.08 (1.02-1.14)
<i>P-value</i>	<i>&lt;0.001</i>	<i>0.005</i>
<b>Semi-automated</b>		
Per 10 bud increase	1.09 (1.04-1.14)	1.06 (1.01-1.11)
<i>P-value</i>	<i>&lt;0.001</i>	<i>0.01</i>

H&E, haematoxylin and eosin; CK, cytokeratin; CI, confidence intervals

## Figure legends

**Figure 1. Methods of assessment of tumour budding.** A-E, manual methods. F-J, semi-automated method. Tumour budding was manually scored on H&E-stained (A, B) and CK-stained (C-J) tissue microarrays. After dearraying, in all cores the tissue border was shrunk by 30 $\mu$ m to exclude candidate buds touching the periphery of the cores. Buds were annotated manually within QuPath (yellow dots in B, D and E). Initial CK buds (CK all, D) were revisited to exclude those lacking a region of nuclear pallor (CK pallor, E) and generate a second dataset applying this criterion. A semi-automated workflow was developed in QuPath (F-J). A binary classifier identified discrete CK positive regions (red). Lumens encapsulated by positive staining were filled in to exclude “luminal pseudobuds” (G). Buds were defined based on area of CK immunopositivity, the acceptable range (40-700  $\mu$ m<sup>2</sup>) derived from analysis of the range of areas of the manually annotated CK buds. Objects with areas outside this range were excluded, leaving buds highlighted (H-J). (H&E, haematoxylin and eosin; CK, cytokeratin)

**Figure 2. Deriving bud area range for semi-automated method.** A, C, E, G and I: Unannotated images. B, D, F, H and J: Corresponding annotated images (red = initial QuPath annotations of CK immunopositivity within broad area range of 5-3000  $\mu$ m<sup>2</sup>; yellow = manual bud annotations). A and B: Tumour gland rupture, generating multiple tiny immunopositive fragments; C and D: Tiny immunopositive fragments of uncertain nature (arrows) detected alongside two true tumour buds (arrowheads); E and F: Six manually annotated (CK all) tumour buds, with areas measured by QuPath (range 107-384  $\mu$ m<sup>2</sup>). G and H: A manually annotated single tumour cell bud with prominent nuclear pallor resulting in underestimation of the bud area by QuPath (measured as 10  $\mu$ m<sup>2</sup>). I and J: Two closely adjacent buds annotated manually (arrows), but considered by QuPath as one large immunopositive area (measured as 820  $\mu$ m<sup>2</sup>). (CK, cytokeratin)

**Figure 3.** A. Histogram of manual CK-detected bud areas, as measured by QuPath. This was the basis of selecting a suitable area range to define a bud applying the semi-automated assessment method. B. Histogram of areas of buds detected by the semi-automated method, applying a range of 40-700  $\mu\text{m}^2$ . (CK, cytokeratin)

**Figure 4. Total bud count comparison across all scoring methods.** A. Total number of tumour buds detected by each of the four methods of assessing tumour budding in all study cases (n=186). B. Total buds per core and frequencies of each number within the study group, for cores with up to 20 buds detected. C. Correlation of total bud counts per core assessed manually by H&E and CK (CK all). D. Correlation of total bud counts per core assessed on CK manually (CK all) and by the semi-automated method. (H&E, haematoxylin and eosin; CK, cytokeratin)

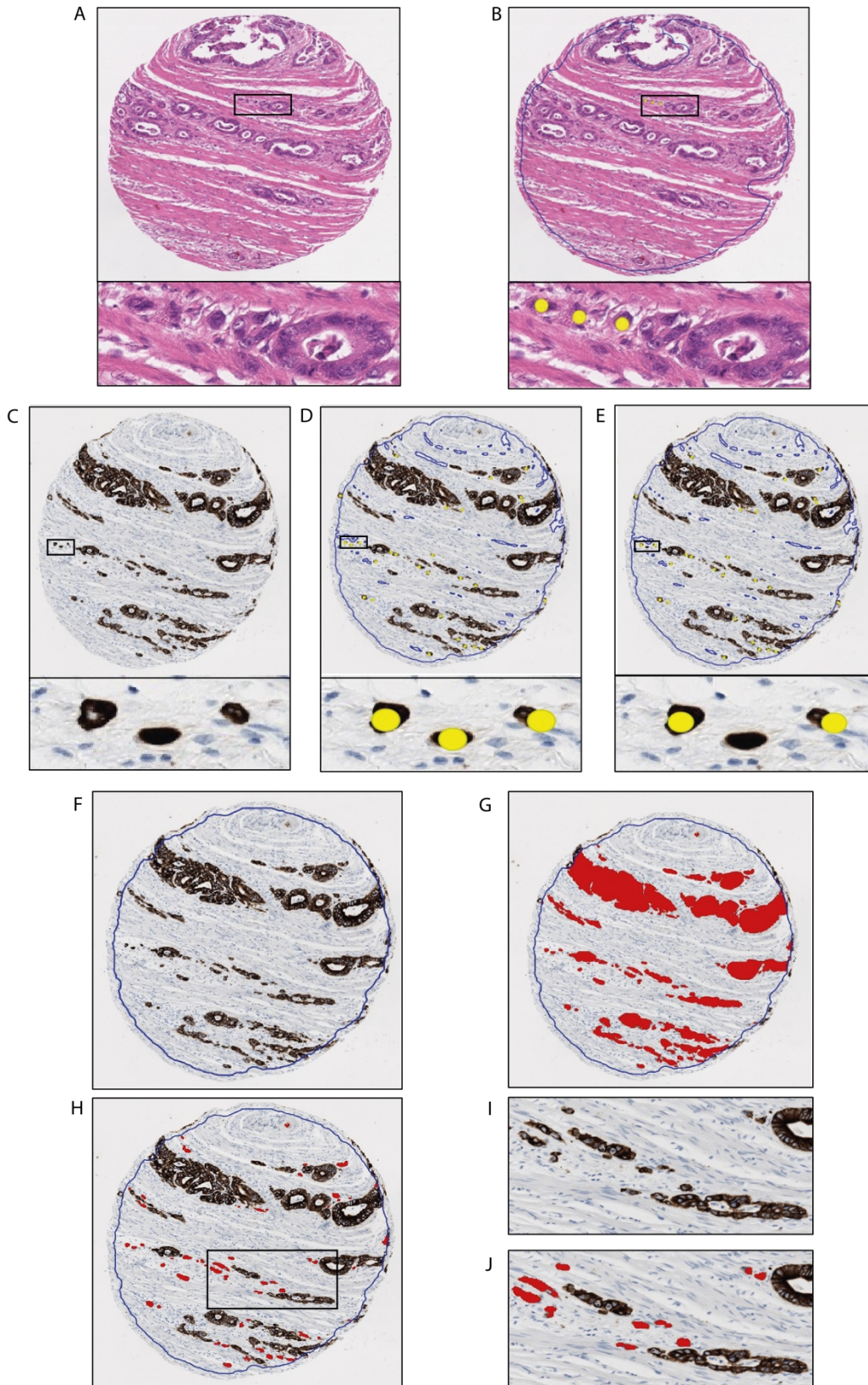
**Figure 5.** Individual bud by bud comparison of manual CK all, manual CK pallor and semi-automated assessment methods. (CK, cytokeratin)

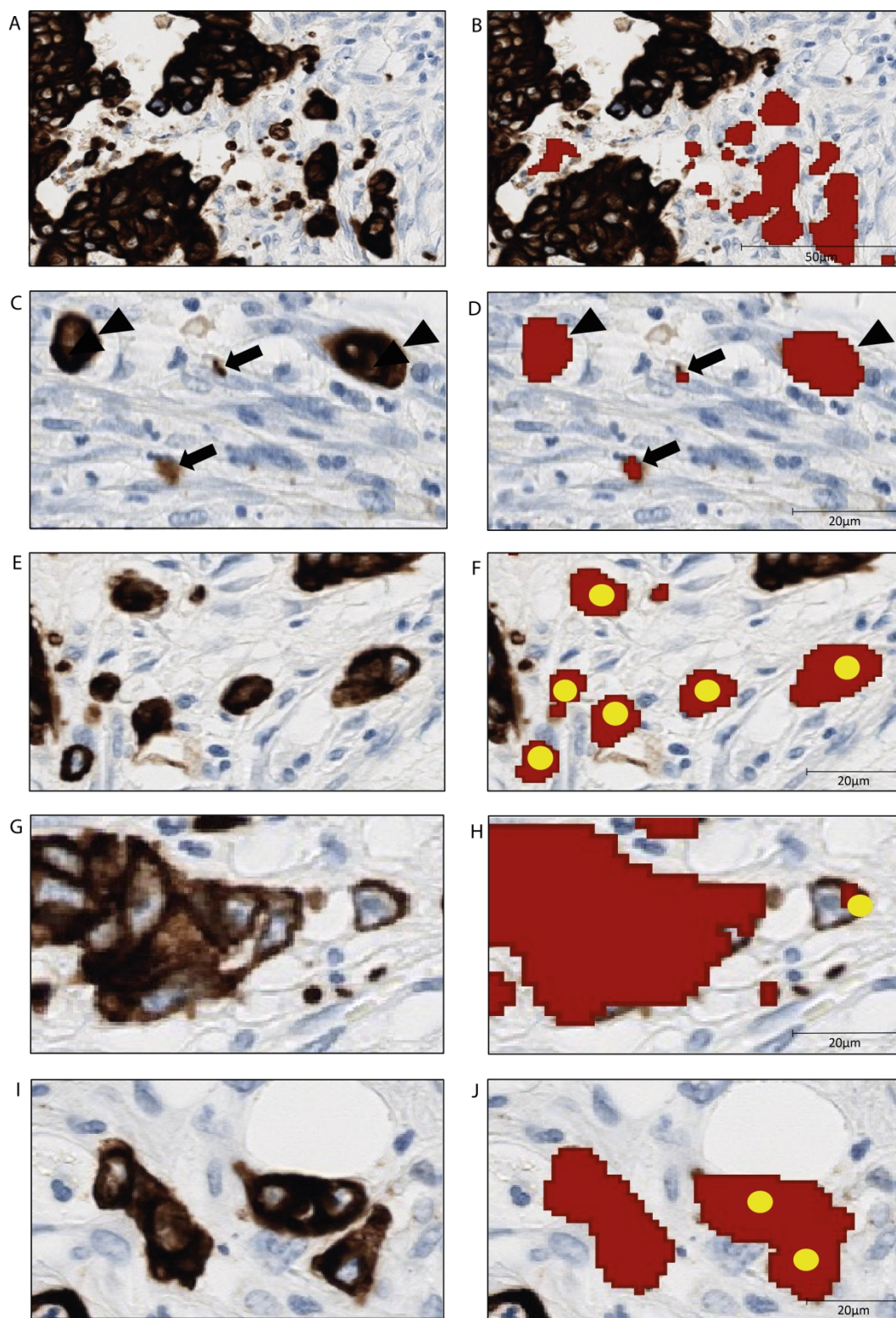
**Figure 6. Discordance in bud assessment between manual CK and semi-automated methods.** A, C, E, G & I: Unannotated images; B, D, F, H & J: Corresponding annotated images (red shapes = QuPath bud annotations; yellow circles = manual bud annotations). A & B: Perfect concordance in annotation of six tumour buds between manual (CK all) and semi-automated methods; C & D: Poor concordance, with the manual method identifying two buds and QuPath identifying two additional, less well-defined, buds; E & F: Mimics of tumour buds within complete glandular lumina ("luminal pseudobuds") are discounted as buds by both manual and semi-automated methods, resulting in concordance; G & H, I & J: If glands are disrupted, resulting in incomplete circumferential immunostaining, QuPath cannot "fill in" the gland lumen and these luminal mimics are counted as buds by the semi-automated method, a particular problem in tissue microarrays when glands involve the core edge (arrows, in I & J). (CK, cytokeratin)

**Figure 7.** Kaplan-Meier estimates demonstrating overall survival differences in patients with stage II/III colon cancer according to low, moderate and high grade tumour budding assessed by four different methods. ITBCC three category cut-offs were utilised for H&E scores ( $\leq 4$ , 5-9,  $\geq 10$  buds) and cut-offs for the other methods scaled up according to the total budding score distribution for each method (Figure 4A). (ITBCC, International Tumour Budding Consensus Conference; H&E, haematoxylin and eosin).

**Figure 8. Tumour budding assessment applied to a whole slide CK-stained image of colorectal cancer.** A high grade budding case has been chosen for illustration. A: After manual annotation of the advancing edge (red line) using the QuPath line tool, the expand annotation tool is used to expand the annotation 1 mm inwards and outwards, delineating the tumour advancing edge region of interest (within yellow outline) for budding assessment. Manually identified (yellow circles) and independently detected QuPath (red shapes) buds are shown (magnified in inset for "hotspot" area); B: Bud density heat map based on manual bud annotations; C: Bud density heat map based on QuPath bud annotations. Density colourmaps are normalised independently for each image according to the range of bud density within the image. The 0.785 mm<sup>2</sup> "hotspot" is highlighted (black circle) in each image. (CK, cytokeratin)

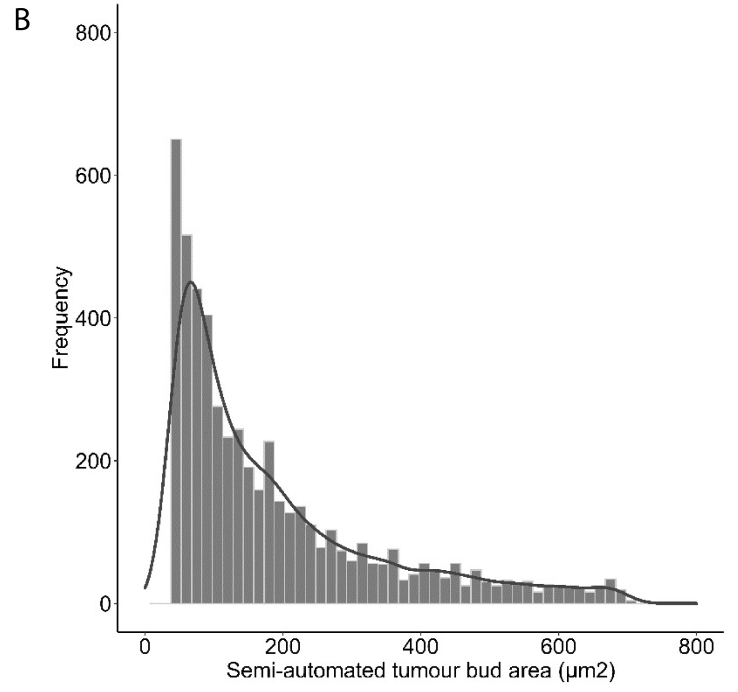
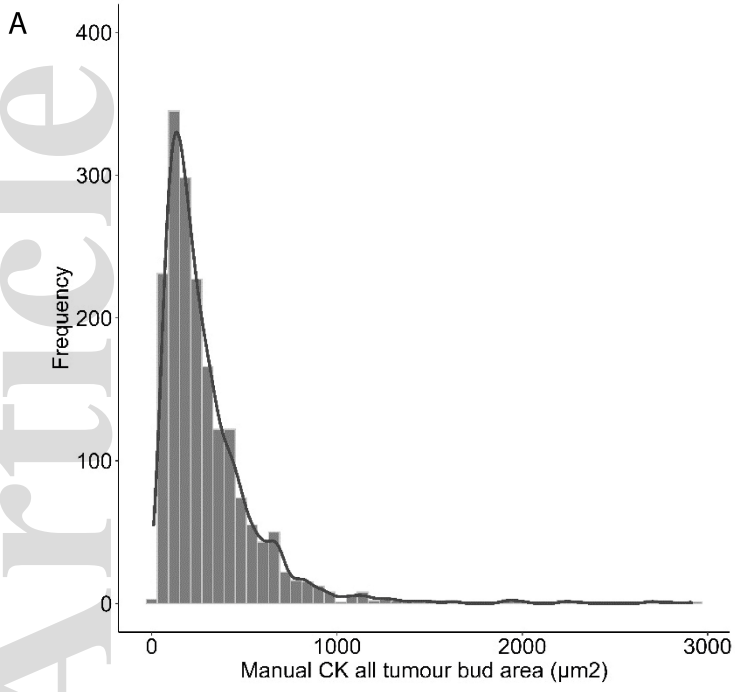
**Supplementary Figure 1.** Representative images of cores which did not meet inclusion criteria. A: Insufficient tumour content; B: Mucinous carcinoma; C: Extensive necrosis present; D: Tumour epithelium with patchy CK immunostaining; E: Significant stromal CK immunopositivity; F: Technical artefact. (CK, cytokeratin)





his\_14574\_f2.jpg



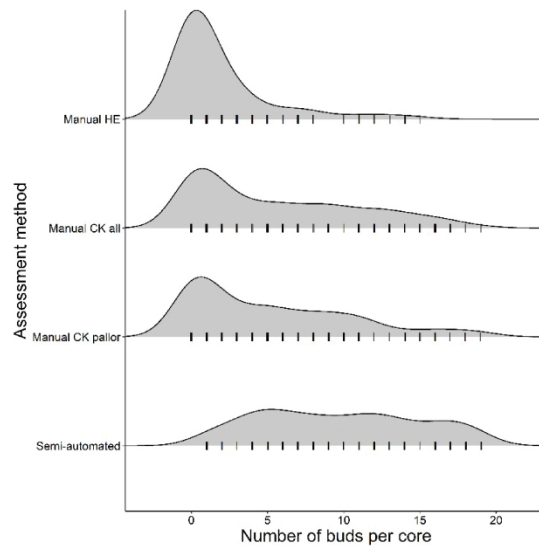


his\_14574\_f3.jpg

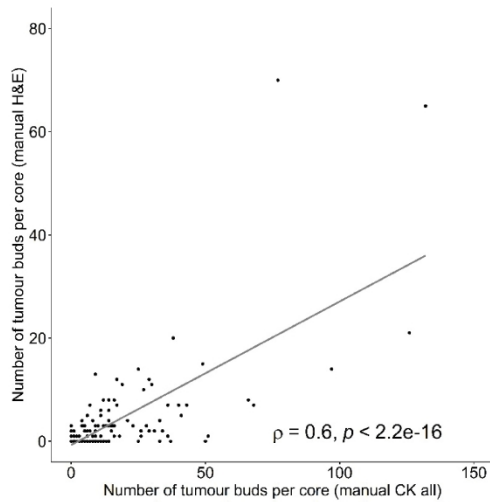
A

Method	Bud definition	Total number of tumour buds
Manual H&E	Single or a cluster of up to four tumour cells on H&E.	503
Manual CK all	Single or a cluster of up to four tumour cells on CK.	2290
Manual CK pallor	Single or a cluster of up to four tumour cells with a visible region of nuclear pallor on CK.	1825
Semi-automated	A region of CK immunopositivity with an area between 40-700 $\mu\text{m}^2$ .	5138

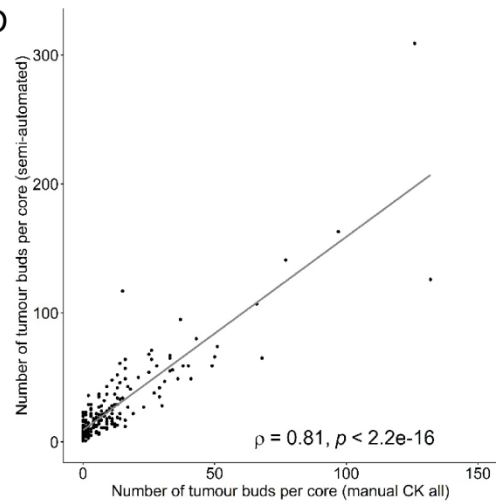
B



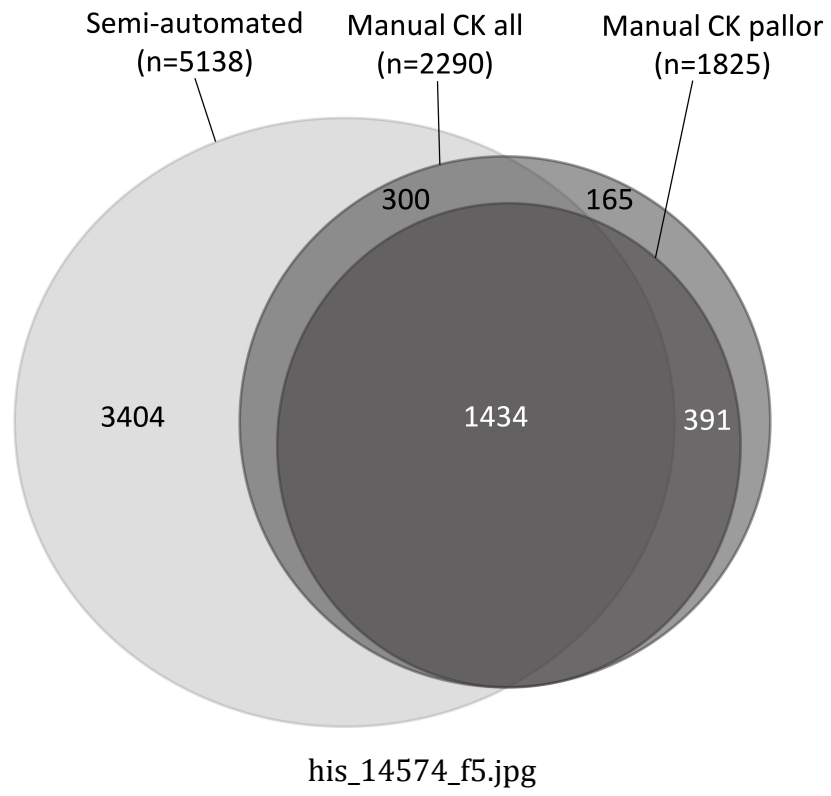
C

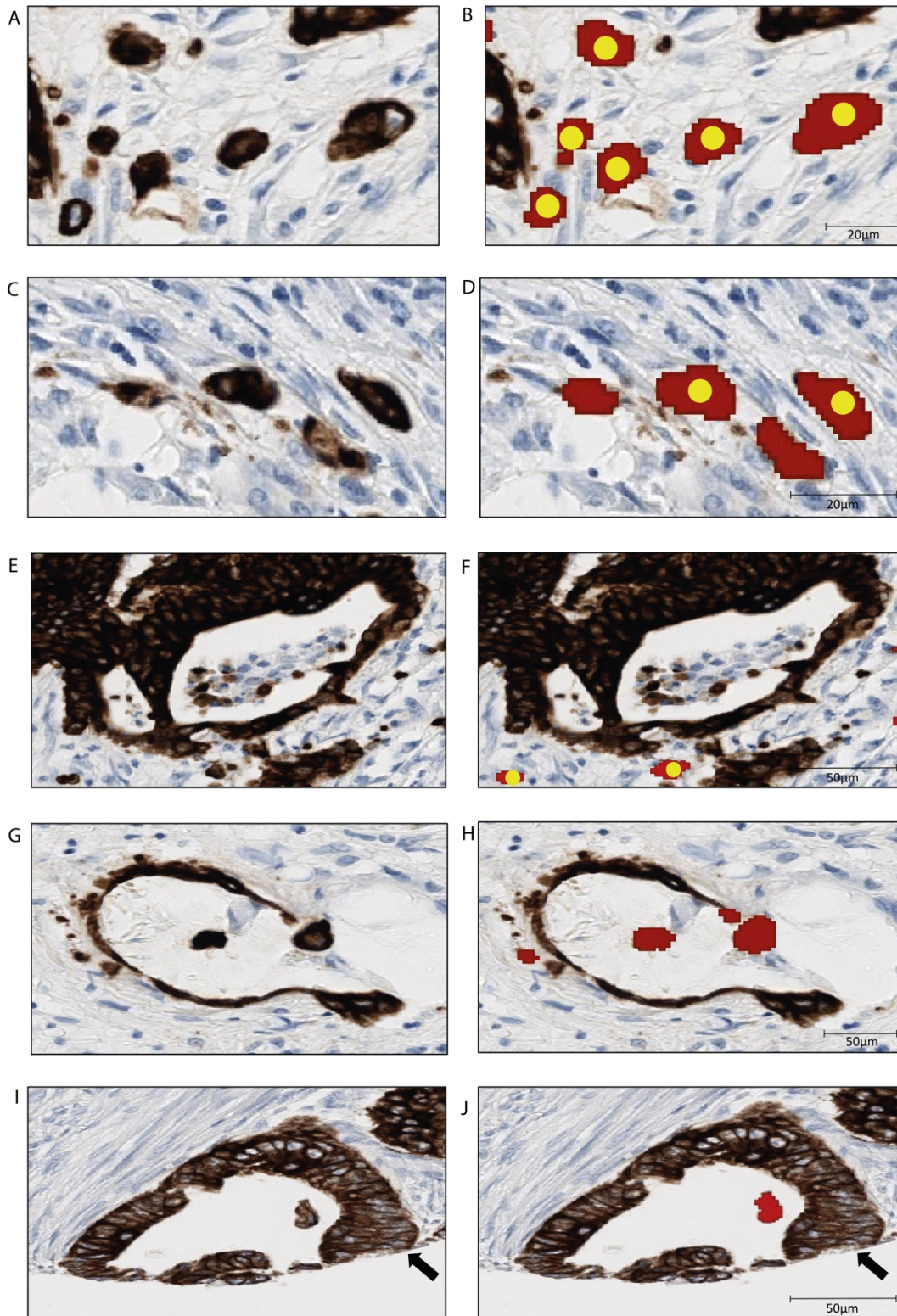


D

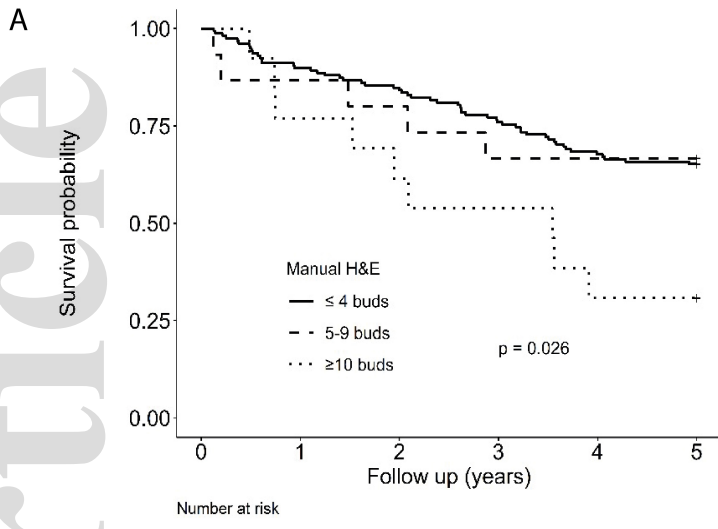


his\_14574\_f4.jpg



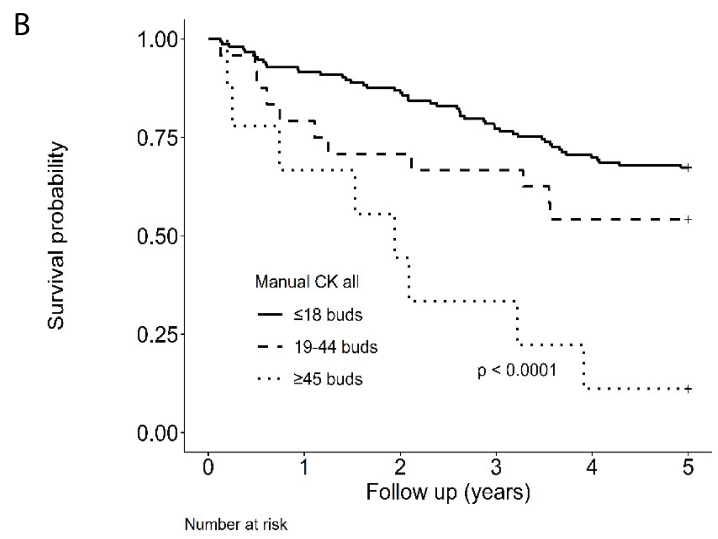


his\_14574\_f6.jpg



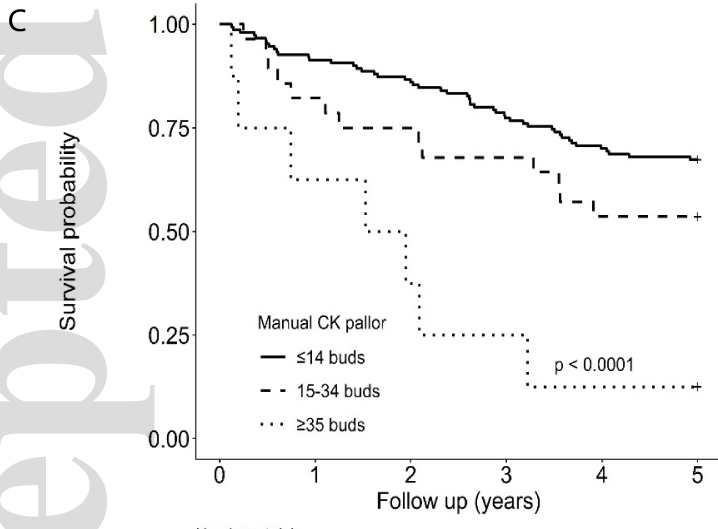
Number at risk

≤ 4 buds	158	142	133	120	107	103
5-9 buds	15	13	12	10	10	10
≥10 buds	13	10	8	7	4	4



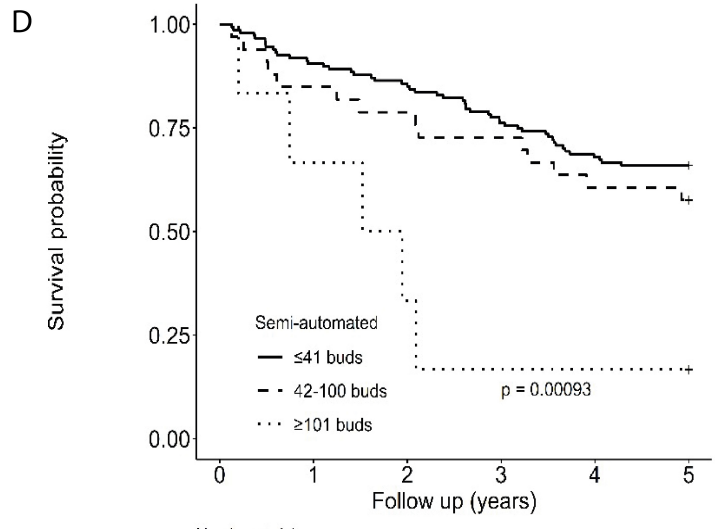
Number at risk

≤18 buds	153	140	132	118	107	103
19-44 buds	24	19	17	16	13	13
≥45 buds	9	6	4	3	1	1



Number at risk

≤14 buds	150	137	129	116	105	101
15-34 buds	28	23	21	19	15	15
≥35 buds	8	5	3	2	1	1



Number at risk

≤41 buds	147	133	125	112	100	97
42-100 buds	33	28	26	24	20	19
≥101 buds	6	4	2	1	1	1

his\_14574\_f7.jpg

