Simultaneous Automatic Scoring and Co-Registration of Hormone Receptors in Tumour Areas in Whole Slide Images of Breast Cancer Tissue Slides

Nicholas Trahearn[a], Yee Wah Tsang[b,c], Ian A Cree[c], David Snead[b,c], David Epstein[d], Nasir Rajpoot[a,c]

[a] Department of Computer Science, University of Warwick, United Kingdom

[b] Department of Pathology & [c] Centre of Excellence for Digital Pathology, University Hospitals Coventry and Warwickshire, United Kingdom

[d] Mathematics Institute, University of Warwick, United Kingdom

Addressee for Correspondence:

Nicholas Trahearn

BioImage Analysis Lab

Department of Computer Science

University of Warwick

Coventry

CV4 7AL

N.Trahearn@warwick.ac.uk

Running Title: Automated Co-Scoring ER/PR Breast Cancer WSIs

Conflicts of Interest:

Nicholas Trahearn is a PhD Student partially funded by Digital Pathology company Omnyx LLC. Omnyx had no further involvement in this study.

ABSTRACT

Aims

Automation of downstream analysis may offer many potential benefits to routine histopathology. One area of interest for automation is in the scoring of multiple immunohistochemical markers in order to predict the patient's response to targeted therapies. Automated serial slide analysis of this kind requires robust registration to identify common tissue regions across sections. We present an automated method for co-localised scoring of Estrogen Receptor and Progesterone Receptor (ER/PR) in breast cancer core biopsies using whole slide images.

Methods and Results

Regions of tumour in a series of fifty consecutive breast core biopsies were identified by annotation on H&E whole slide images. Sequentially cut immunohistochemical stained sections were scored manually, before being digitally scanned and then exported into JPEG 2000 format. A two-stage registration process was performed to identify the annotated regions of interest in the immunohistochemistry sections, which were then scored using the Allred system. Overall correlation between manual and automated scoring for ER and PR was 0.944 and 0.883 respectively, with 90% of ER and 80% of PR scores within in one point or less of agreement.

Conclusions

This proof of principle study indicates slide registration can be used as a basis for automation of the downstream analysis for clinically relevant biomarkers in the majority of cases. The approach is likely to be improved by implantation of safeguarding analysis steps post registration.

Key Terms: Automation, Breast, Immunohistochemistry, Estrogen Receptors, Progesterone Receptors, Molecular Pathology

Introduction

The adoption of digital whole slide imaging for the analysis of histopathology slides (often referred to as digital pathology) presents novel opportunities for improving and increasing efficiency of the way pathologists work[1]. The increasing disease burden presented by the elderly population and escalating costs of treatment place growing demands on many clinical services to provide more for less, improvements in efficiency may prove to be key in achieving this goal.

Immunohistochemical (IHC) staining of oestrogen receptor (ER) and progesterone receptor (PR) is an essential part of the pathological assessment of breast carcinomas. Scoring is a visually estimated quantification of the reactivity of a given IHC marker, using qualities such as stain intensity and proportion. Scoring of these markers is known to predict response to hormonal treatment.[2-4] ER and PR are both nuclear markers with similar expression properties, and thus the scoring process for both markers is identical. Visual scoring by a pathologist is subjective and therefore prone to both inter- and intra-observer variation[5]. It is also something that can consume much of a pathologist's time, as they are required to re-examine the case after staining in order to provide the score. There have been a number of studies[6-8] into automated scoring of these markers, however analysis has typically been restricted to single stain analysis, and supervised by the pathologist. Although helpful in improving consistency, such studies do not markedly reduce pathologist's workload. Reliable co-localisation of regions of interest, that are known to contain tumour, across several sequentially cut sections theoretically allow for an almost entirely automated analysis of IHC staining expressions for multiple markers at a time. The pathologist would need only to identify the region of the tumour on the initial diagnostic section, and the equivalent regions of interest for the other sequentially cut sections could then be located and analysed automatically. This would thereby avoid the need for the case to be re-examined by the pathologist.

In histopathology, the need to section the sample removes the spatial relationships between serial sections along the third dimension. This is significant when we wish to make multiple observations on the same sub-region of the tissue, such as when studying the expression profile of several IHC stains on a localised region of interest (ROI). As such, we require a robust method of bringing sections from different slides back into a common alignment, such that we can identify the same region of tissue across many sections. This is a process known as registration. Registration of this kind is approximately rigid, that is to say that we can align the two sections closely with just translation and rotation. This, however, does not consider the small physical distortions to the tissue that naturally occur during the sectioning process. While these are unlikely to have a noticeable impact on the scale of a whole section, small offsets from the rigid alignment are to be expected at a local level.

We present a simultaneous ER/PR automated scoring method, built upon a two-stage registration process that aligns serial tissue sections. This allows the system to co-locate the same region of interest for multiple sections from the same sample, and thus provide like-for-like scores for corresponding tumour areas in both sections. The automatic ER/PR scoring follows the protocols outlined in the Allred Scoring system,[9] a scoring system used in regular clinical practice. We show that our method achieves good correlation with the pathologist's own scores on a common set of breast cases.

Materials and Methods

Ethics, Tissue Samples, Staining and Image Acquisition

Ethical approval for this study was granted by the National Research Ethics Service - Dulwich Committee 12/LO/0993.

A series of 50 consecutive breast core biopsies positive for primary breast carcinoma were used for this study. Once the diagnosis had been made sequential sections were cut at a thickness of 5 microns: For each case, a Haematoxylin and Eosin (H&E) section was taken and stained using the Sakura Tissue-Tek Prisma staining machine (Sakura Finetek Europe B.V. KvK / Chamber of Commerce Leiden 28065449). In addition, a pair of serial sections were stained with oestrogen receptor ER(SP1) and progesterone receptor PR(1E2) antibodies, respectively, using an automated Ventana BenchMark immunostaining machine. Staining was carried out according to the manufacturer's instructions and visualized using the manufacturer's recommended 3,3' diaminobenzidine (DAB) visualization kit Ventana iVIEWTM DAB (Ventana Medical Systems Inc Roche Group, 1910 Innovation park, Tuscon AZ85755 USA). Antibodies were commercially supplied pre-diluted and pre-treated by Ventana Medical Systems Inc. Positive control tissue from previously stained breast cancer cases was included on the test section. Negative controls were performed on sequential sections using the same staining method with the exception of the primary antibody. Sections were consistently taken in the following order:

1. H&E
2. ER
3. PR

The cases included in this study were selected before the full adoption of digital pathology, consequently the reporting pathologist scored the ER/PR stained sections on glass, using the Allred scoring system. All the slides were then digitised using the Omnyx VL120 (Omnyx LLC 1251 Waterfront Place, Pittsburgh PA15222 USA) slide scanner. Slides were scanned with a resolution of 0.275μm/pixel, to produce the equivalent of a 40× magnification. Regions of tumour were then annotated on the digitised H&E section by a second pathologist, who had not viewed the case before and did not view the immunohistochemistry. Images were then converted from a proprietary format into JPEG 2000 format and exported for analysis by the automated algorithm.

Section Registration:

It is not possible to automatically identify a common scoring area on both the ER and PR sections from a given case

without first re-aligning the sections with each other. Consequently, automated registration of the sections was required before scoring could take place. The section registration was coarsely a two-stage process: an initial approximate alignment, followed by a series of small corrective alignments. The approximate alignment was made on the basis of the external boundary of the section, while the corrective alignments were made by matching common structures of the registered sections, such as fat and clusters of nuclei. The flow of the registration algorithm is shown as a block diagram in Figure 1.

Approximate Registration

We initially aligned the boundaries of tissue sections by matching their points of locally maximal curvature.[10] Curvature is a feature that can vary depending on the resolution, or the level of detail, at which the curve is observed, which is illustrated in Figure 2. Curvature Scale Space (CSS)[11] is a representation of a polygon's curvature across a set of decreasing resolutions, simulated by smoothing the polygon with Gaussian filters of increasing sigma. Figure 3 demonstrates the maxima of curvature found at different resolutions.

We used a CSS representation of the section's boundary to identify points of locally maximal curvature at different levels of detail. It is likely that for similar shapes, such as the boundaries of serial tissue sections, the points of the curvature maxima will be at similar locations on each boundary for all but the highest resolutions. Thus, it should be possible to find a close match between the sets of curvature maxima from the two boundaries. Finding that match will allow us to generate the best-fit rigid transformation for registration. Matching between maxima can be thought of as an assignment problem and was solved using the Hungarian algorithm.[12] It should be noted that this approach may produce many equally good registration candidates if the sections external boundary is rotationally symmetric, such as for circular sections commonly used on tissue microarrays. This, however, was not the case for the diagnostic breast core biopsies used within our study.

Registration Refinement

The approximate registration step described above produced the initial alignment. This alignment is a rigid transformation or rather, a transformation composed only of translation and rotation. A further refinement step was then applied on top of the initial alignment, to produce the final alignment. This step was intended to correct instances where the approximate transformation was incorrect at a local level. The registration refinement was performed again each time that the field of view (FOV) is changed, to ensure that the best alignment was always presented to the scoring

algorithm. This method aligned the tissue sections based upon their inner structure, isolating tissue structures, such as glands and nuclei clusters, and attempting to find correspondence between them.

We assume that our initial alignment is correct at a low resolution, and thus is already close to the optimal alignment. The refinement registration step should therefore produce a transformation that is close to our initial alignment.

In contrast to the initial alignment, the optimal alignment is likely to be non-rigid and non-linear. However, this is only the case on the scale of a whole section and, for smaller FOVs at a high resolution, the non-linearities are likely to be minor and difficult to distinguish. Therefore, for the type of FOVs used in scoring, the corrective alignment was modelled as a rigid transformation that varies slightly from our initial alignment.

This design decision has two key advantages. First, rigid alignments are less complex and are therefore faster to compute. Secondly, non-linear transformations have the potential to distort the tissue, perhaps in a way that does not accurately reflect the biological structure of the tissue. This is not true of rigid transformations, which preserve structure.

Automatic ER/PR Scoring

The automated Allred Scoring was performed using a newly developed algorithm. For each case, the ER/PR sections were co-registered to a H&E slide in the case. The H&E slide contained an annotation, specifying a region of tumour to be scored by the algorithm, necessitating registration between the H&E section and the ER/PR section. Consequently, the ER/PR sections were only scored by the algorithm in the tissue regions corresponding to the areas annotated by the pathologist on the case's corresponding H&E slide.

The complete pipeline of the algorithm, outlined in Figure 4, was modelled on the Allred Scoring System.[9] This system categorises the IHC stain intensities into four groups: unstained, weak, intermediate, and strong and assesses the proportion of neoplastic nuclei stained. It is important to note that, while there are guidelines for these groups, the scoring is ultimately at the discretion of the pathologist. As such, there are no strict definitions for the boundaries between the groups. For this work we defined the thresholds for the IHC stain intensity groups as shown in Figure 5. The values for the boundaries were determined empirically by sampling pixels from control tissue sections belonging to

each stain group.

Only the staining on nuclei within the tissue were scored. In addition, for ER/PR scoring, we are only interested in the staining that occurs to tumour nuclei. On occasion the IHC stain binds to non-tumour nuclei,[13-15] but this has no relevance to the score and should be ignored. Therefore, to ensure that only the correct parts of the tissue are being scored, the tumour nuclei must be isolated from the remaining tissue. To achieve this we employed two detectors, a tumour region detector and a nuclei detector. Each detector produced a binary mask and the intersection of the two masks yielded our desired result, the tumour nuclei regions.

In certain cases, the human visual system may identify a region as purely strongly stained, but a closer inspection of the visual field shows a much wider variety of stain intensities. In some cases, this can interfere with the scoring process, as the stain group with the numerical majority may not necessarily appear as the dominant stain. Therefore, rather than using the DAB channel directly, we first pre-processed the channel to reduce the variation of the stain intensity within each nucleus. Figure 6 demonstrates the reduction of the intensity variation for strongly stained nuclei as a result of this pre-processing step.

Following extraction of the ER/PR stained tumour nuclear regions, the proportion of the nuclei that belong to each of the four aforementioned stain categories was calculated by extracting all tumour nuclei pixels and binning them into the four categories according to their intensities in the processed IHC stain channel. The number of pixels in each bin was recorded and the values were then normalised such that their sum was 1, which thereby produced our estimate for the proportions of nuclei in each stain category.

The two parts of the Allred score were computed from the estimated proportions of each stain intensity, using the protocol as defined by the College of American Pathologists,[9] which is outlined in Tables 1a and 1b. The term "positively stained" refers to nuclei that are either weakly, intermediately, or strongly stained, and thus was calculated as the sum of those three respective proportions.

One minor alteration to the protocol was the introduction of a minimum proportion threshold for positively stained nuclei. If the proportion of positively stained nuclei was estimated as less than 0.01%, then the field of view was automatically scored as 0. This measure was introduced to address the possibility of a small collection of pixels being

falsely assessed as stained. In most cases these pixels had no influence on the result because of their low frequency, however in negative cases there are no other stained pixels and thus their influence was far more pronounced. For our chosen FOV parameters, 0.01% of the total nuclei pixels would amount to an area much smaller than a single nucleus, and thus the addition of this threshold was unlikely to cause underscoring.

The automated scoring process was performed independently on twenty random, non-overlapping visual fields at a resolution equivalent to a 20x magnification. The modal score was taken as the final score for the annotated region. If two or more scores were equally popular then a further five visual fields were selected and scored, this process was repeated until the deadlock was resolved. For validation, the automated scoring result was compared with a manual score from a single pathologist on the same section. The pathologist performing the manual scoring was not the same as the pathologist providing the tumour region annotation.

Results

Automated and manual scoring was performed on a set of 50 patient cases, each containing one slide for ER and one slide for PR, using the processes outlined above. Neither the pathologist performing manual scoring nor the individual running the automated scoring algorithm were given access to the scores from their experimental counterpart until all scoring was finalised. Figure 7 shows the initial results of the pathologist's scoring compared to that of the automated approach. In all 50 cases the registration and scoring algorithm successfully identified and scored the ROI identified by the pathologist.

In 80% of ER slides and 54% of PR slides, the algorithm scored exactly the same as the pathologist. In addition, for 90% of ER slides and 78% of PR slides the algorithm's score was within one of the pathologist's score. For this experiment Allred scores of 2 were considered as being within one of a score of 0. The mean absolute error of the automatic scoring was found to be 0.40 for the ER sections and 0.92 for the PR sections, resulting in a mean error of 0.66 across both ER and PR sections. For the purpose of these calculations scores of 0 were changed to 1 in order to ensure equal numerical spacing between the possible scores.

The correlation between pathologist's scores and automated scores was 0.881, demonstrating significance to a p value of <0.001. Correlation for the scoring of just ER sections was 0.922, with a p value of <0.001. Correlation for the scoring of just PR sections was 0.840, with a p value of <0.001.

In the initial scoring results, a number of discrepancies were observed between pathologist's and algorithm's scores, in some cases resulting in a different diagnostic outcome. Scoring differences that were diagnostically significant were subsequently recorded and flagged for further review. Reviewed slides were sent to a different pathologist for rescoring. To prevent potential bias, the pathologist performing rescoring was not made aware of the scores produced by the algorithm and original pathologist. In addition, the algorithm's FOV scores were checked to identify any potential issues in the automated scoring. A summary of the inconsistent slides is presented in Table 2, with a potential root cause of the scoring disagreement listed alongside each slide.

Figure 8 shows the results of scoring following review. In 80% of ER slides and 56% of PR slides, the algorithm scored exactly the same as the pathologist. For 90% of ER slides and 80% of PR slides, the algorithm's score was within one of the pathologist's score. The mean absolute error of the automatic scoring was found to be 0.36 for the ER sections and

0.80 for the PR sections, resulting in a mean error of 0.58 across both ER and PR sections.

The correlation between pathologist's scores and automated scores was 0.914, demonstrating significance to a p value of <0.001. Correlation for the scoring of just ER sections was 0.944, with a p value of <0.001. Correlation for the scoring of just PR sections was 0.883, with a p value of <0.001.

The kappa agreement measure for the pathologist and algorithm scores was 0.601 for ER slides, 0.464 for PR slides, and 0.535 for the combined set of slides. Caution should be taken when interpreting basic kappa statistics for Allred scoring, as all disagreements are treated equally in the calculation. Clearly this is not the case, as certain disagreements in scoring are likely result in different therapies, for instance Allred scores of 0 and 8, whereas for other disagreements this is far less likely, such as Allred scores of 7 and 8. Therefore, we also calculated the weighted kappa statistics, where larger and potentially diagnostically significant disagreements were weighted more heavily, as shown below:

$$W_{i,j} = 1 \quad \text{if } |P_i - A_j| > 1$$

$$W_{i,j} = 0.5 \quad \text{if } |P_i - A_j| = 1$$

$$W_{i,j} = 0 \quad \text{if } |P_i - A_j| = 0$$

Where $W_{i,j}$ is the weight on the disagreement between pathologist score P_i and algorithm score A_j. For the purpose of calculation Allred scores of 0 were converted to 1 to ensure uniform spacing of the scoring range. Under this weighting we calculated kappa agreement of 0.753 for ER slides, 0.676 for PR slides, and 0.7177 for the combined set of slides.

The entire scoring process typically takes around 2 seconds per slide, which includes the time taken for registration.

Discussion

In this study, we have demonstrated a two stage system for co-registration of sequential cut histopathology sections, and illustrated how such an approach can be used to provide automation of both the co-localisation of regions of interest and the scoring of ER and PR in these regions. We have shown in another work[16] that the system is capable of high quality registration of multi-IHC and is able to update the alignment when necessary, such as when the initial approximate alignment has not produced the best registration at a local level.

The automated system showed a high level of agreement with the pathologist, comparable with that of the inter-observer agreement between two pathologists. A correlation of 0.85 for ER and 0.87 for PR has been shown for the agreement between pathologists on a set of 74 resected breast cancer specimens[17]. Pathologist error rates as high as 24% have also been reported[18]. One study found pathologist vs. pathologist kappa agreements of 0.57 for ER and 0.51 for PR, however this value dropped to 0.42 for ER and 0.36 for PR when comparing pathologists from different laboratories[19].

The use of algorithms to prevent scoring in benign background epithelium as opposed to tumour was successful in avoiding false positive scoring. The pathologist is still required to identify the tumour area to be scored as a region of interest in order to distinguish invasive carcinoma from in-situ disease. One problem of this approach is likely to occur in cases exhibiting heterogeneous staining, where areas of the tumour selected may, for a variety of reasons, stain very differently to other areas of the tumour. An example of heterogeneous staining is shown in Figure 9a. This may result in mis-scores by the algorithm in situations where the initial region selected by the pathologist is not representative of the entire section, two examples of which were found in our data. Errors of this type could potentially be prevented by the addition of a fail-safe algorithm, which would perform tumour segmentation outside of the marked region of interest and then assess whether the staining and score were consistent with the marked region. We would recommend that cases failing the verification process for staining homogeneity be referred for further manual analysis.

Section thickness is an important consideration in serial section registration, IHC scoring, or any other cross-slide analysis task. Thicker sections are less likely to have common tissue structures across many serial sections, and it will therefore be more difficult to score the exact same tissue region for each marker as parts of the region may not be present on other sections. It may also be more difficult to perform registration if the method relies on common tissue structures, such as is the case for our registration refinement method. For our data, it was found that the chosen section

thickness of 5μm does not appear to cause any issues of this kind.

In two cases, over-scoring was seen in negative (pathologist score 0). These were the result of poor stain separation, which resulted in the system overestimating the quantity of IHC staining on the section. These two cases are the result of the algorithm selecting fields of view with minimal staining of any kind, which could potentially be addressed by introducing further restrictions on the visual fields used to estimate the Allred score.

One large source of errors arose from scoring the raw IHC channel, for slides that were given a score of 8 by the pathologist. These were often scored as 7 by the algorithm. This was because the algorithm was scoring the slide as a 2 for intensity, instead of 3. The reason for this is due to the way the human eye observes intensity. Often the regions that are visually discerned as strongly stained are actually composed of pixels with many different stain intensities, as shown in Figure 6. In such cases, despite the clear presence of a strong stain intensity, pixels with intermediate staining may in fact be numerically more frequent than the strongly stained pixels. In other words, the stain group with the numerically predominant intensity is not always the visually predominant intensity. This problem can be addressed by adopting the previously discussed filtering steps on the IHC channel, which were designed to increase the intensity of stain pixels that are surrounded by strong staining. It may prove to be the case that precise pixel stain intensities give a better prediction of response to steroid receptor targeted therapy than the conventional semi-quantitative visual analysis on relative intensity. This, however, is beyond the scope of this study, which is focused on emulation of the established methods of human visual assessment.

For four slides the algorithm was able to identify possible faults in the original pathologist's scores, as shown by the notable improvement of pathologist-algorithm scoring agreement following a review of discrepancies in the initial scoring results. This in itself can be seen as a potential example of the use of automated scoring for the purposes of verification of manual scores. In a diagnostic workflow where manual scoring is preferred, such a scoring algorithm could still potentially be used as a means of identifying possible mis-scores. In situations where the pathologist and algorithm's scores do not agree, such a case could then automatically be assigned to an additional pathologist for further analysis.

The three remaining discrepant slides, listed in Table 2, were all instances of poor slide quality. For the slide with an out-of-focus WSI, the blurring caused the DAB stain to appear much less intense, as can be seen in Figure 9b. Thus, the

algorithm estimated Allred intensity score to be a lower value than if the image had been in focus. This was purely an issue with the slide's scanning and subsequent digitisation, and thus was simply resolved by rescanning the slide and performing automated scoring on the new WSI.

The DAB stain's failure to bind to an area of this tissue on one slide had produced an artefactual shadow of negative staining, which is shown in Figure 9c. The presence of this shadow had the effect of lowering the estimated proportion scores, and thus produced a lower overall Allred score.

The dark artefacts, shown on the section in Figure 9d, were caused during the slide preparation when drying the back of the coverslip. The dark regions are, in some cases, falsely identified as IHC staining, which resulted in an overestimation of the proportion of positive nuclei, and thus the proportion score.

The aforementioned slide quality issues highlight a clear need for accurate quality assurance algorithms within any future automated diagnostic workflow. This will allow unsuitable WSIs to be identified and replaced at an early stage, thereby preventing possible false calls from downstream algorithms. Automated detection of out-of-focus regions on a slide is one potential quality assurance algorithm that could be employed prior to automated scoring.

In conclusion, we have shown that automated registration of regions of interest identified by the reporting pathologist, and the assessment of ER and PR is feasible. Our findings indicate that, in a laboratory using WSIs for reporting of breast cancer cases, the pathologist need only examine the slide once at the time of diagnosis. The post-diagnosis analysis of IHC markers can be devolved to image analysis systems, that are able to locate a common region of interest by registration of sequentially cut sections. The strategy has the scope to improve both consistency of scoring and efficient use of healthcare resources.

Acknowledgements

David Snead, Ian Cree, and Nasir Rajpoot designed the study. David Snead and Yee Wah Tsang collected image data and annotated tumour regions. Nicholas Trahearn created the automated algorithms under supervision of Nasir Rajpoot. David Epstein contributed to the development of the approximate registration algorithm. Nicholas Trahearn performed the automated scoring experiments. Nicholas Trahearn, David Snead, Ian Cree, Yee Wah Tsang, and Nasir Rajpoot wrote the paper. All authors were involved in the editing of the paper. Whole-slide images used during the course of this project have been acquired using Omnyx scanners installed at University Hospitals Coventry and Warwickshire. Research has been funded jointly by EPSRC and Omnyx LLC.

List of Abbreviations

ER: Oestrogen Receptor

PR: Progesterone Receptor

DAB: 3,3'-Diaminobenzidine

H&E: Haematoxylin and Eosin

ROI: Region of Interest

FOV: Field of View

IHC: Immunohistochemical

CSS: Curvature Scale Space

RGB: Red, Green, and Blue

WSI: Whole Slide Image

Supporting Material

A: Description of Image Analysis Algorithms

B: Description of Multi-Stain Analyser Software

References

1. Snead DR, Tsang YW, Meskiri A et al. Validation of digital pathology imaging for primary histopathological diagnosis. Histopathology 2016; 68: 1063-72.
2. Viale G, Regan MM, Maiorano E et al. Prognostic and predictive value of centrally reviewed expression of estrogen and progesterone receptors in a randomized trial comparing letrozole and tamoxifen adjuvant therapy for postmenopausal early breast cancer: BIG 1-98. J Clin Oncol 2007; 25: 3846-3852.
3. Fisher B, Dignam D, Bryant J et al. Five versus more than five years of tamoxifen therapy for breast cancer patients with negative lymph nodes and estrogen receptor-positive tumors. J Natl Cancer Inst 1996; 88: 1529-1542.
4. Putti TC, El-Rehim DMA, Emad AR et al. Estrogen receptor-negative breast carcinomas: a review of morphology and immunophenotypical analysis. Mod Pathol 2005; 18: 26-35.
5. May M. A better lens on disease. Sci Am 2010; 302: 74-77.
6. Ali HR, Irwin M, Morris L et al. Astronomical algorithms for automated analysis of tissue protein expression in breast cancer. Br J Cancer 2013; 108: 602-612.
7. Bolton KL, Garcia-Closas M, Pfeiffer RM et al. Assessment of automated image analysis of breast cancer tissue microarrays for epidemiologic studies. Cancer Epidemiol Biomarkers Prev 2010; 19: 992-999.
8. Khan AM, Mohammed AF, Al-Hajiri SA et al. A novel system for scoring of hormone receptors in breast cancer histopathology slides. IEEE Middle East Conference on Biomedical Engineering 2014; 155-158.
9. Lester SC, Bose S, Chen YY, Connolly JL. Protocol for the examination of specimens from patients with invasive carcinoma of the breast. Arch Pathol Lab Med 2009; 133: 1515-1538.
10. Trahearn N, Epstein D, Snead D, Cree I, Rajpoot N. A fast method for approximate registration of whole-slide images of serial sections using local curvature. Proc SPIE Int Soc Opt Eng Medical Imaging 2014; 90410E-90410E.
11. Mokhtarian F. Silhouette-based isolated object recognition through curvature scale space. IEEE Trans. Pattern Anal Mach Intell 1995; 17: 539-544.
12. Kuhn HW. The hungarian method for the assignment problem. Naval Research Logistics Quarterly 1955; 2: 83-97.
13. Umekita Y, Yoshida H. Immunohistochemical study of hormone receptor and hormone-regulated protein expression in phyllodes tumour: comparison with fibroadenoma. Virchows Arch 1998; 433: 311-314.

14. Sapino A, Bosco M, Cassoni P et al. Estrogen receptor-β is expressed in stromal cells of fibroadenoma and phyllodes tumors of the breast. Mod Pathol 2006; 19: 599-606.

15. Knower KC, Chand AL, Eriksson N et al. Distinct nuclear receptor expression in stroma adjacent to breast tumors. Breast Cancer Res Treat 2013; 142: 211-223.

16. Trahearn N, Epstein D, Cree I, Snead D, Rajpoot N. Hyper-Stain Inspector: A Framework for Robust Registration and Localised Co-Expression Analysis of Multiple Whole-Slide Images of Serial Histology Sections. Submitted.

17. Cohen DA, Dabbs DJ, Cooper KL et al. Interobserver agreement among pathologists for semiquantitative hormone receptor scoring in breast carcinoma. Am J Clin Pathol 2012; 138: 796-802.

18. Tuominen VJ, Ruotoistenmäki S, Viitanen A et al. ImmunoRatio: a publicly available web application for quantitative image analysis of estrogen receptor (ER), progesterone receptor (PR), and Ki-67. Breast Cancer Res 2010; 12: 1-12.

19. Regitnig P, Reiner A, Dinges H-P et al. Quality assurance for detection of estrogen and progesterone receptors by immunohistochemistry in Austrian pathology laboratories. Virchows Arch 2002; 441: 328-334.

20. Fox D, Burgard W, Thrun S, Cremers AB. Position estimation for mobile robots in dynamic environments. Proc Conf AAAI Artif Intell 1998; 983-988.

21. Myronenko A, Song X. Point set registration: Coherent point drift. IEEE Trans Pattern Anal Mach Intell 2010; 32: 2262-2275.

22. Schubert W, Bonnekoh B, Pommer AJ et al. Analyzing proteome topology and function by automated multidimensional fluorescence microscopy. Nat Biotechnol 2006; 24: 1270-1278.

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

24. Macenko M, Niethammer M, Marron JS et al. A method for normalizing histology slides for quantitative analysis. Proc IEEE Int Symp Biomed Imaging 2009; 9: 1107-1110.

25. Khan AM, Rajpoot N, Treanor D, Magee D. A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. IEEE Trans Biomed Eng 2014; 61: 1729-1738.

26. Hyvärinen A, Oja E. Independent component analysis: algorithms and applications. Neural Netw 2000; 13: 411-430.

27. Trahearn N, Snead D, Cree I, Rajpoot N. Multi-class stain separation using independent component analysis. Proc SPIE Int Soc Opt Eng Medical Imaging 2015; 94200J-94200J.

Predominant Stain Intensity	Intensity Score
No Stain	0
Mostly Weakly Stained	1
Mostly Intermediately Stained	2
Mostly Strongly Stained	3

Table 1a: A summary of the intensity scoring criteria for the Allred scoring system.

Percentage of Positively Stained[Note 1] Nuclei	Proportion Score
0% – ≤0.1%	0
0.1% – ≤1%	1
1% – ≤10%	2
10% – ≤33%	3
33% – ≤66%	4
66% – 100%	5

Table 1b: A summary of the proportion scoring criteria used for the Allred scoring system.

[Note 1] Positively stained nuclei refer to all IHC stained tumour nuclei, regardless of their intensity.

Case	Stain	Notes
U20	ER	Heterogeneous staining.
U42	PR	Heterogeneous staining.
U05	ER	Overestimated DAB staining due to poor stain separation.
U16	PR	Overestimated DAB staining due to poor stain separation.
U03	PR	Disagreement in scores by original and reviewing pathologist.
U45	ER	Disagreement in scores by original and reviewing pathologist.
U45	PR	Disagreement in scores by original and reviewing pathologist.
U48	PR	Disagreement in scores by original and reviewing pathologist.
U15	PR	Slide is out of focus. Algorithm re-scored on re-scanned slide.
U12	PR	Artefactual shadow on slide.
U01	PR	Overestimated DAB staining due to dark artefacts on slide.

Table 2: A summary of the slides with large Pathologist-Algorithm scoring disagreements.

Figure Legends

Figure 1: A block diagram showing the overall flow of the automated registration and scoring procedure. Please refer to Figure 4 for further details on the automated ER/PR scoring algorithm.

Figure 2: A demonstration of curvature as a multi-resolution feature. The curve segment shown has two types of curvature maxima: the high resolution bumps and the low resolution turn across the entire curve, each maxima is highlighted by a red circle.

Figure 3: A demonstration of curvature maxima of the given tissue boundary extracted at different resolutions, each shown with a red circle. Lower resolution curves are generated by Gaussian smoothing the original. Top-left shows the original section image.

Figure 4: A block diagram showing the details of the automated ER/PR scoring algorithm.

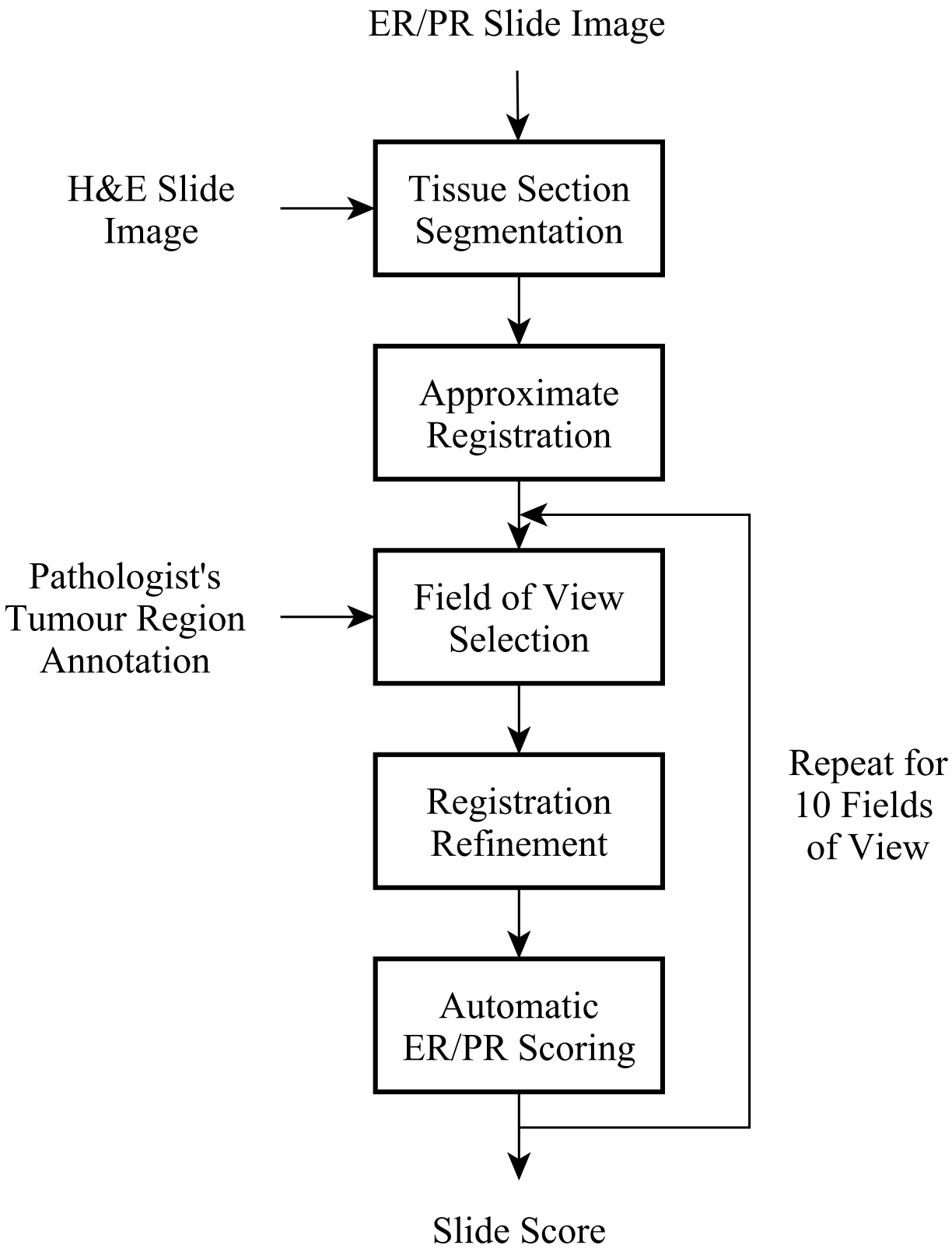
Figure 5: An illustration of the boundaries between each stain intensity group, used by the automated scoring algorithm. Locations of the boundaries were determined empirically using control tissue samples.

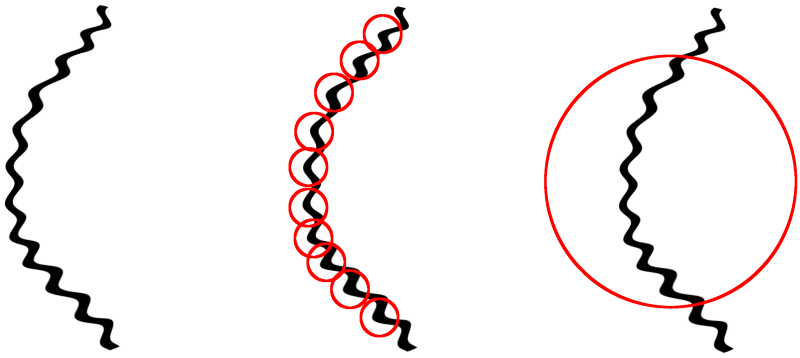
Figure 6: A diagram demonstrating the intensity variation of strongly stained nuclei, both before and after a specialised filtering operation. The filtering operation is intended to reduce the amount of variation of staining within each nucleus.

Figure 7: Scatter plots showing the results of the automatic ER/PR scoring, compared to the pathologist's manual scores. Size of the dot is proportional to the number of cases with the given Pathologist-Algorithm score pair, the number of cases is also placed alongside the dot. All results between the solid green lines have the same score as the pathologist. All results between the dotted red lines are within one of the pathologist’s score.

Figure 8: Scatter plots showing the results of the automatic ER/PR scoring compared to the pathologist's manual scores following re-scoring of discrepant sections. Size of the dot is proportional to the number of cases with the given Pathologist-Algorithm score pair, the number of cases is also placed alongside the dot. All results between the solid green lines have the same score as the pathologist. All results between the dotted red lines are within one of the pathologist’s score.

Figure 9: Examples of potential sources of scoring discrepancies between the pathologist and algorithm. a) Heterogeneous staining. b) Out of focus WSI. c) Artefectual shadow. d) Dark artefacts caused by drying back of coverslip.

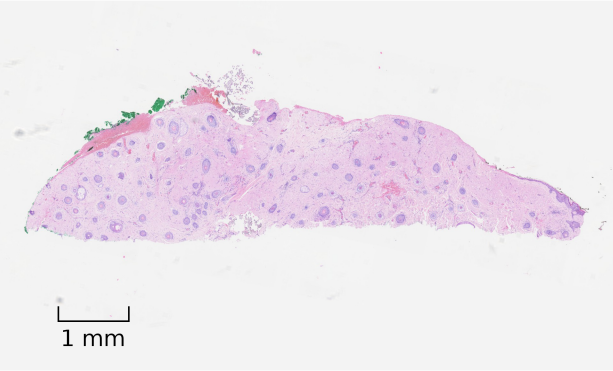

ER/PR Slide Image
H&E Slide Image
Tissue Section Segmentation
Approximate Registration
Pathologist's Tumour Region Annotation
Field of View Selection
Registration Refinement
Automatic ER/PR Scoring
Repeat for 10 Fields of View
Slide Score

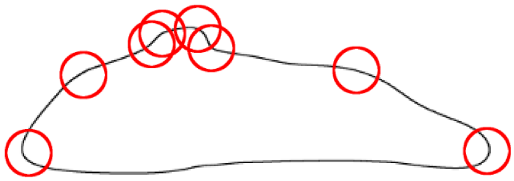


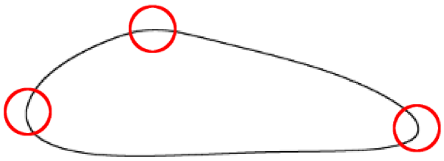
Line Segment

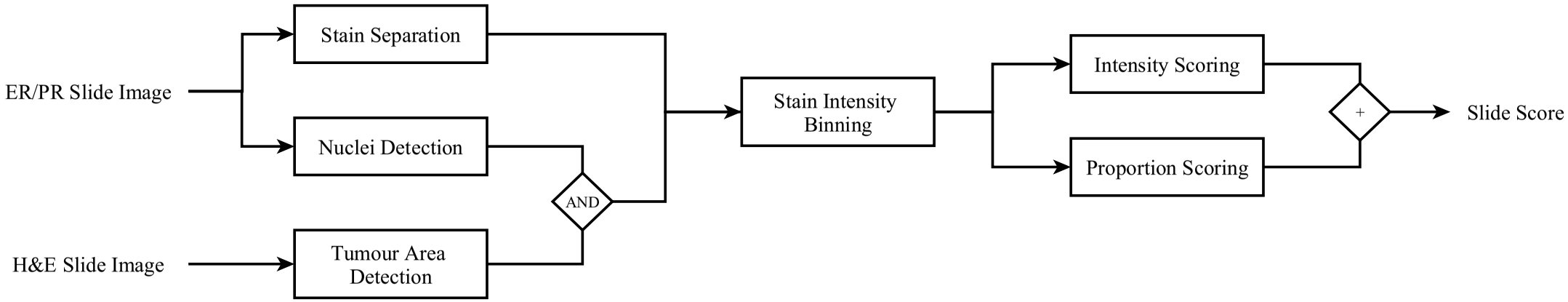
High Resolution
Curvature Maxima

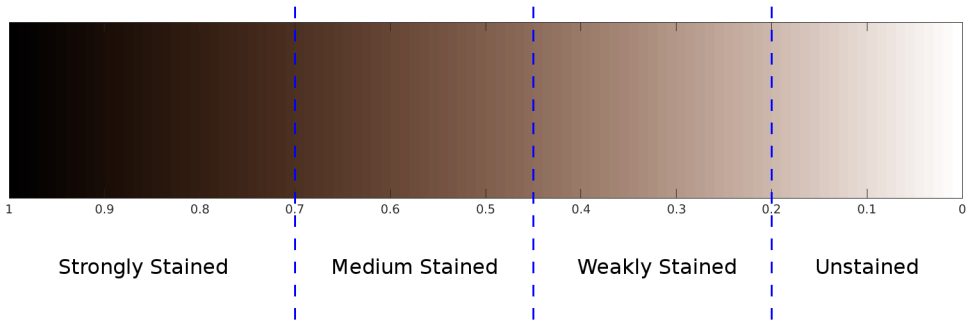
Low Resolution
Curvature Maxima

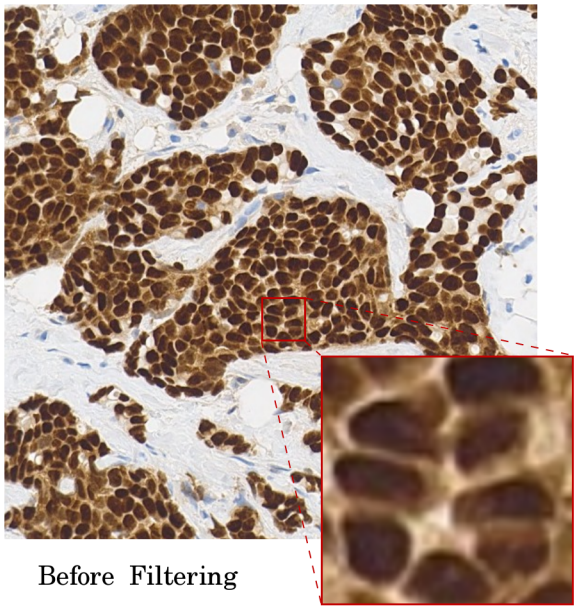

1 mm



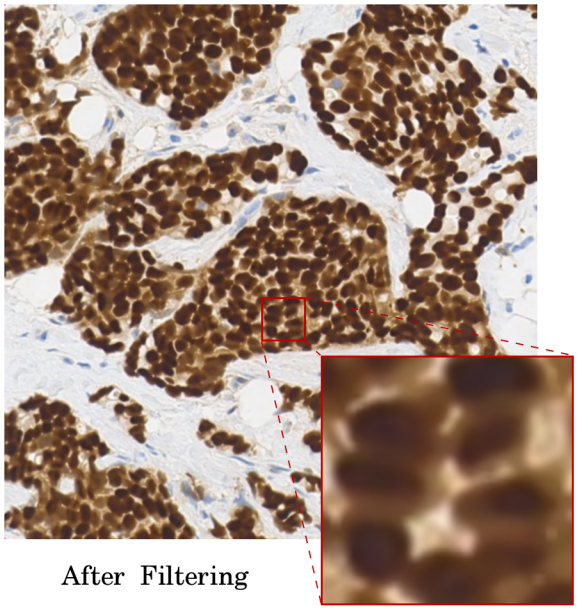



ER/PR Slide Image
H&E Slide Image
Stain Separation
Nuclei Detection
Tumour Area Detection
AND
Stain Intensity Binning
Intensity Scoring
Proportion Scoring
+
Slide Score

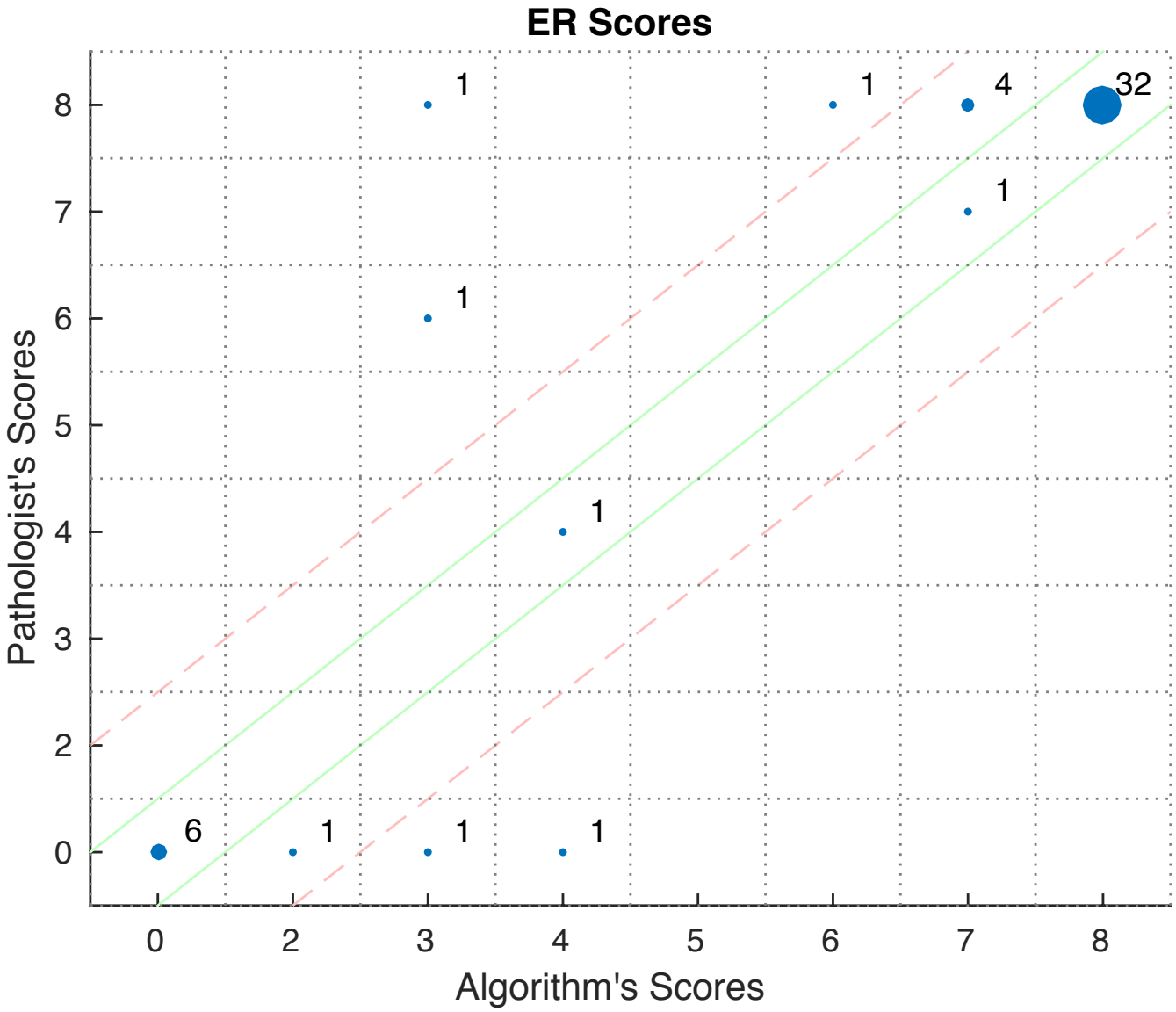

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Strongly Stained
Medium Stained
Weakly Stained
Unstained

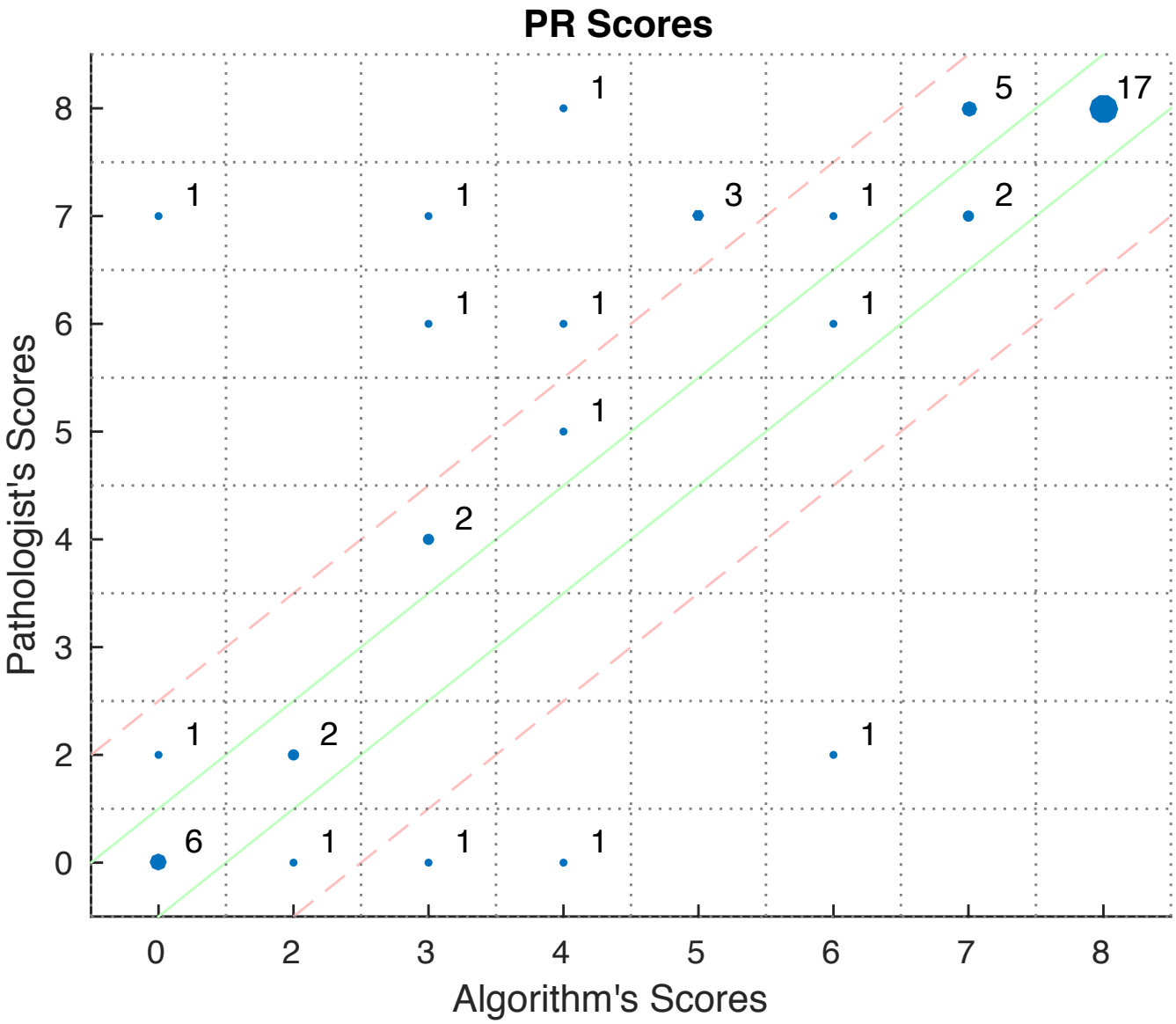


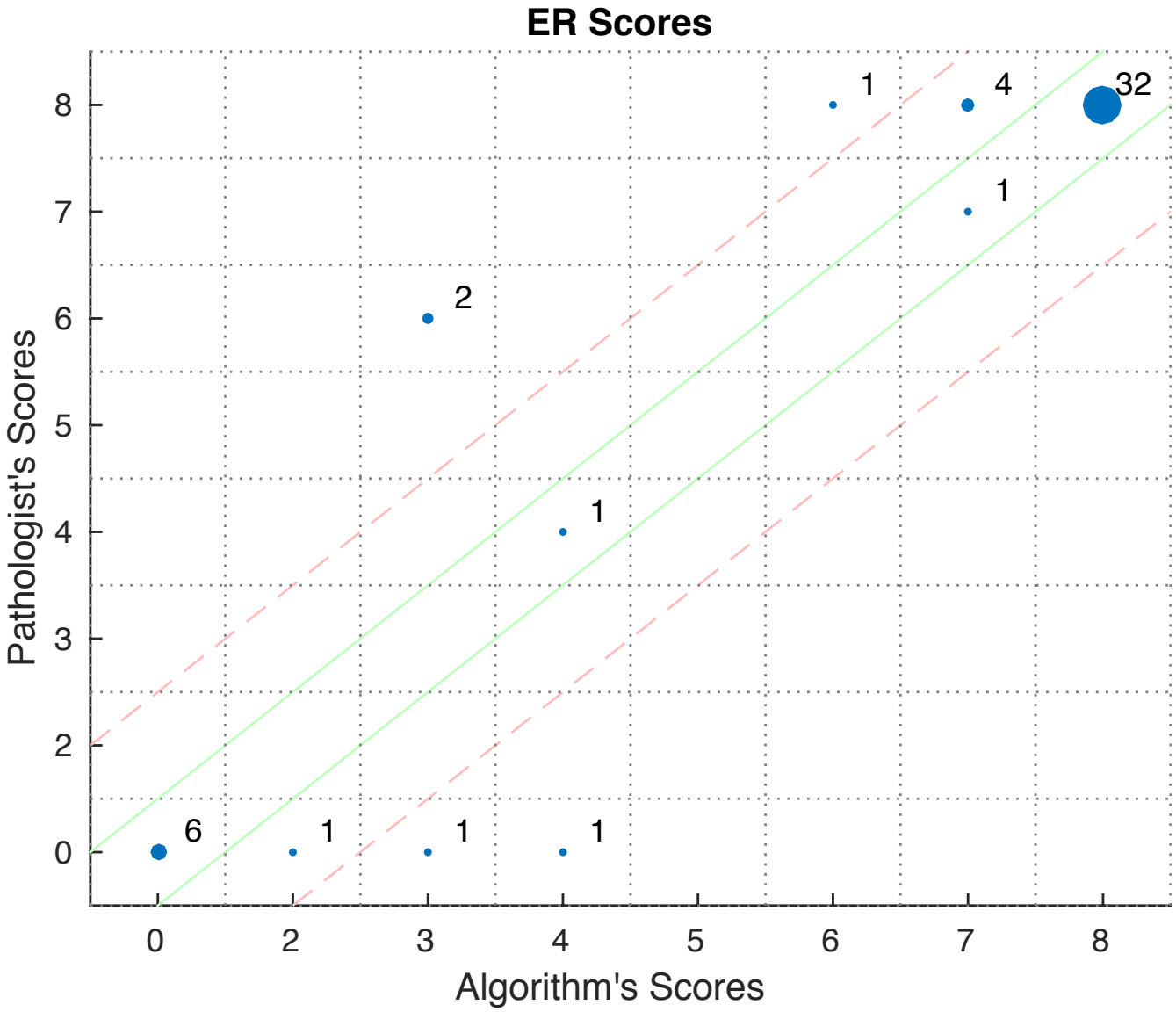
Before Filtering

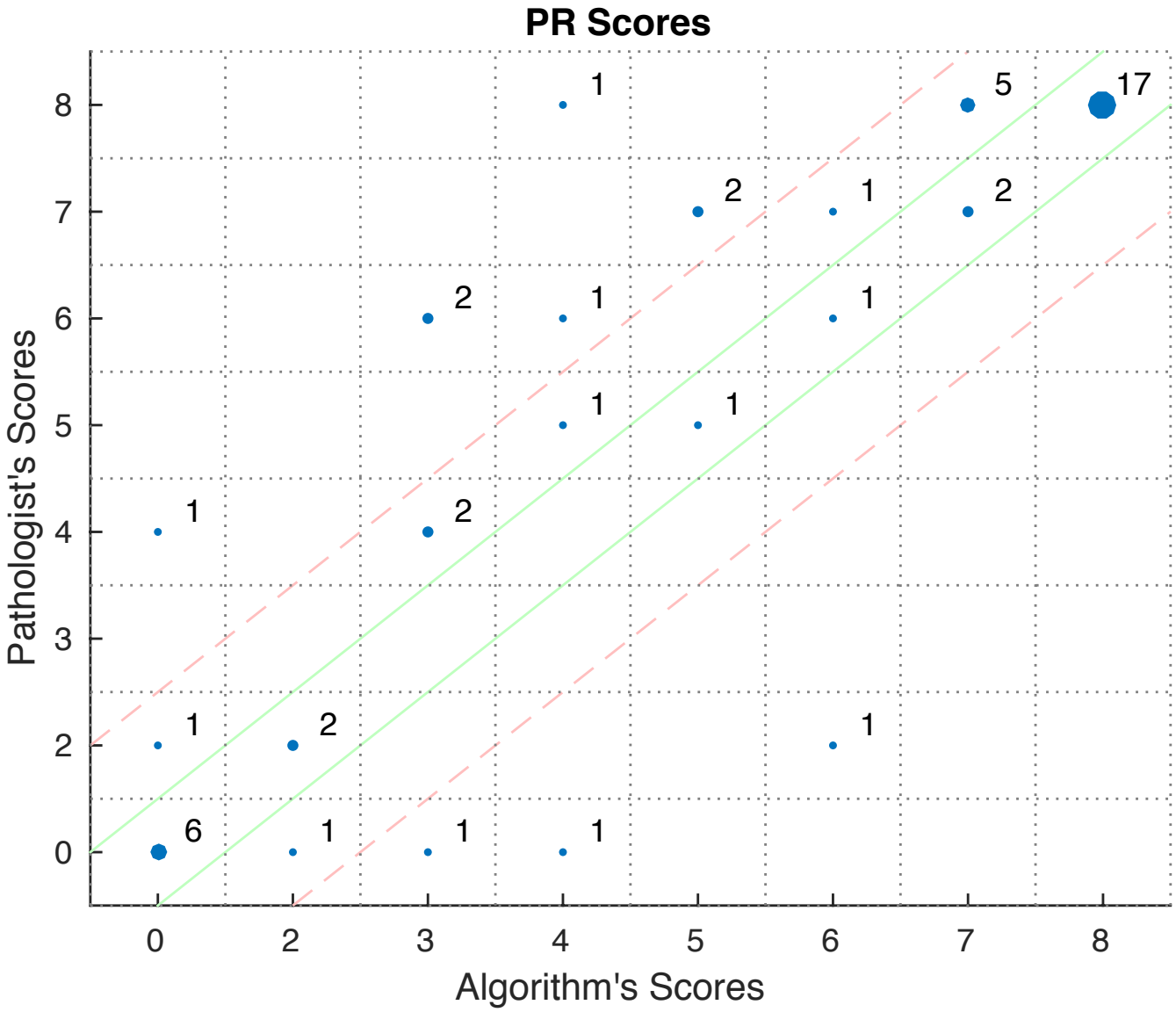


After Filtering


ER Scores
Pathologist's Scores
Algorithm's Scores
0
2
3
4
5
6
7
8
1
1
4
32
1
1
1
6
1
1
1


PR Scores
Pathologist's Scores
Algorithm's Scores
0
2
3
4
5
6
7
8
1
5
17
1
1
3
1
2
1
1
1
1
1
2
1
2
1
6
1
1
1


ER Scores
Pathologist's Scores
Algorithm's Scores
0
2
3
4
5
6
7
8
1
4
32
1
2
1
6
1
1
1


PR Scores
Pathologist's Scores
Algorithm's Scores
0
2
3
4
5
6
7
8
1
5
17
2
1
2
2
1
1
1
1
1
2
1
2
1
6
1
1
1