A Systematic Review and Meta-analysis of Automated Tools for HER2, ER and T-cell Scoring in Cancer Biopsies

Anna-Maria Tsakiroglou[1,2], Susan Astley[3,4], Kim Linton[1,2,5], Anne Martel[6], Isabel Peset-Martin[7], Catharine West[1,5], Richard Byers[1,8], and Martin Fergie[3]

[1] Division of Cancer Sciences, University of Manchester, Manchester Academic Health Science Centre, Manchester, UK; [2] Manchester Cancer Research Centre, Manchester, UK; [3] Division of Informatics, Imaging and Data Sciences, School of Health Sciences, University of Manchester, Manchester, UK; [4] Prevent Breast Cancer and Nightingale Breast Screening Centre, Wythenshawe Hospital, Manchester University NHS Foundation Trust, Manchester, UK; [5] The Christie NHS Foundation Trust, Manchester, UK; 6 Department of Medical Biophysics, University of Toronto, Sunnybrook Research Institute, Toronto, Canada; [7] Discovery Science & Technology, Medicines Discovery Catapult, Alderley Park, UK; [8] Manchester Royal Infirmary, Manchester University NHS Foundation Trust (MFT), Manchester, UK.

Abstract

Background: Computer assisted scoring (CAS) in pathology should improve standardisation, reproducibility, and throughput. This review assesses the validation approaches and performance of CAS tools developed for HER2, ER and T-cell scoring in tumour tissue to provide a baseline to judge new algorithms and identify attributes needed for clinical adoption.

Methods: PubMed, Web of Science, and IEEE Xplore Digital Library were queried to retrieve peer-reviewed studies quantitatively validating CAS tools. Study quality was assessed using standardised criteria (protocol CRD42019139688). The algorithms and validation approaches were described. Agreement of CAS with pathologists was assessed in random effects meta-analysis; studies reporting Cohens κ of agreement against manual ER Allred scoring and manual HER2 scoring (0/1+, 2+, 3+) were included.

Results: Moderately good agreement with manual scoring was observed for HER2 (κ = 0.75, 95% CI: 0.70-0.81) and ER algorithms (κ = 0.74, 95% CI: 0.66-0.83). Automated and pathologist generated scores agreed at least as well as those between pathologists for HER2 and T-cells, but not ER. Multiple pathologists providing annotations improved performance for HER2. CAS increased inter-observer agreement compared to manual scoring, however, it did not reduce the need for confirmatory HER2 fluorescent in-situ hybridization testing.

Conclusions: Validation of CAS should not only demonstrate agreement with pathologists, but also a benefit over conventional scoring practice, by improving reproducibility, robustness to staining variability or correlation to patient outcome. As study heterogeneity and lack of context description can hinder adoption, we suggest there is a need for reporting guidelines for validating of such tools.

Introduction

Significant advances in slide scanning technology pave the way for a digital pathology revolution. Alongside digitisation, computer assisted scoring (CAS) tools are being developed to support the pathologists workflow. These tools aim to quantify the number and intensity of stained objects in tissue images, and offer a promising avenue towards better standardisation, reproducibility and throughput. However, it is unclear whether they are sufficiently accurate and reliable to warrant routine clinical adoption, and there are no established best practices for reporting their validation.

This review focuses on CAS for the expression of nuclear oestrogen receptor (ER), membranous human epidermal growth factor receptor-2 (HER2) and cytoplasmic T-cell markers (cluster of differentiation [CD] 3, CD4, CD8). While ER and HER2 expression levels are assessed routinely in breast cancer to select adjuvant therapies, T-cell populations in the tumour micro-environment are increasingly studied for their impact on cancer prognosis.

This review aimed to investigate the performance of CAS for ER, HER2 and T-cell markers in cancer biopsies. Studies validating automated scoring algorithm were retrieved and the approach used for their validation was described. We report a meta-analysis of the agreement between automated and pathologist generated scoring. The objectives were to: i) provide a baseline to judge the performance of new algorithms and ii) outline their effectiveness on attributes fundamental for clinical adoption such as the inter-observer variability.

Methods

Technology, study design and image preparation: CAS tools were reviewed for HER2, ER and T-cell markers (CD3, CD8, CD4). Peer-reviewed studies, providing quantitative validation were included. Selection criteria based on a pilot screening process and agreed by author consensus were: formalin-fixed, paraffin-embedded (FFPE) or frozen tissue of human tumours or adjacent stroma, published after 1/1/2000 with full text available. Any staining protocol, detection system and scanning setup was included, however, this information was recorded to provide context. Studies were excluded if they involved animal tissue, blood, cellular aspirates, cell lines or bony tissue.

Comparators and outcomes: The performance of CAS is judged against comparators including human pathologists scoring or alternative methods for measuring antibody quantity. The type of comparators and performance measures were recorded. However, studies validating algorithms against multiple markers without reporting performance for each individual marker were excluded.

Synthesis of results: Conceptual mapping describes the algorithm components of CAS systems. The approach for validation, types of reference ground truth and performance metrics were summarised by narrative synthesis. Potential sources of heterogeneity in the validation setup for different markers were explored. The performance of automated scoring systems was probed by quantitative meta-analysis of Cohens κ metric. Cohen's κ measures inter-rater variability, assessing whether the degree of agreement between two alternative forms of a test (here automated algorithm vs pathologist) is higher than expected by chance. Studies were included if they reported data required for calculating Cohens κ and its variance, as described by Sun [2]. For HER2, most studies reported the Cohens κ using the 3-tier ACSOC/CAP HER2 scoring system (0/1+, 2+, 3+). For ER, most studies reported Cohens κ for the dichotomised Allred score (≤ 2 = negative, > 2 = positive). The limited number of T-cell studies prohibited quantitative meta-analysis.

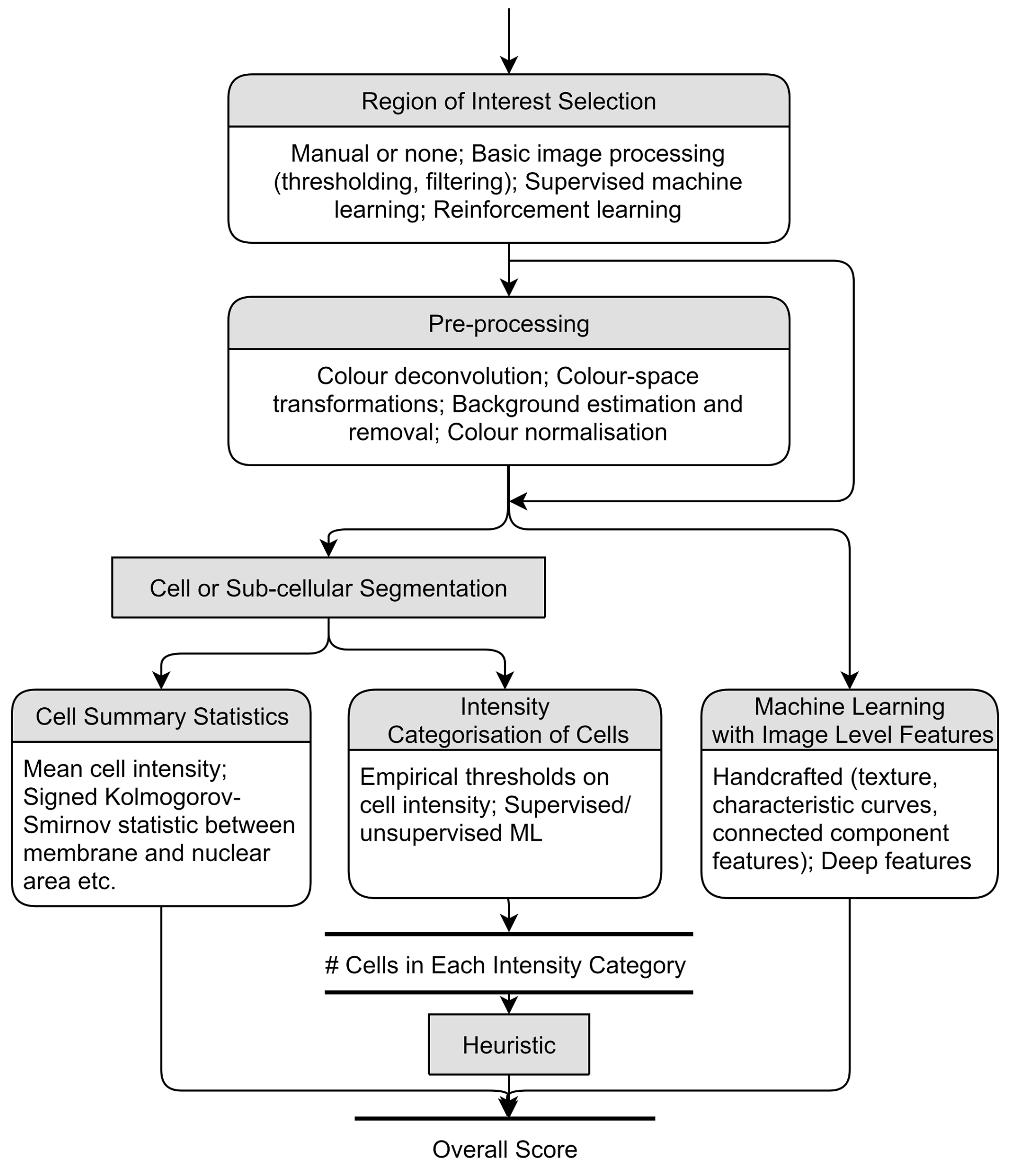
Region of Interest Selection
Manual or none; Basic image processing (thresholding, filtering); Supervised machine learning; Reinforcement learning
Pre-processing
Colour deconvolution; Colour-space transformations; Background estimation and removal; Colour normalisation
Cell or Sub-cellular Segmentation
Cell Summary Statistics
Mean cell intensity; Signed Kolmogorov-Smirnov statistic between membrane and nuclear area etc.
Intensity Categorisation of Cells
Empirical thresholds on cell intensity; Supervised/ unsupervised ML
Machine Learning with Image Level Features
Handcrafted (texture, characteristic curves, connected component features); Deep features
# Cells in Each Intensity Category
Heuristic
Overall Score

Fig. 1: Conceptual map of common scoring algorithm structures.

Results

Studies included in review: Ninety-six studies were identified for qualitative synthesis and thirteen for quantitative meta-analysis. Several studies validated more than one algorithm; 65 algorithm validations are reported for HER2, 49 for ER and 13 for T-cell markers. For quantitative meta-analysis, nine HER2 studies (11 comparisons) and five ER studies (6 comparisons) provided sufficient information to calculate the Cohens κ against pathologists ground truth. Frequent reasons for exclusion were lack of automated methodology or quantitative validation, and markers other than those predefined.

Description of scoring algorithms: Figure 1 summarises the common approaches for automated scoring algorithms.

HER2 Agreement with pathologists: Random effects meta-analysis of HER2 studies placed the summary estimate of Cohens κ at 0.75 (95% CI: 0.70-0.81). However, high heterogeneity was present with Higgins I^2=79.5% (95% CI: 54.2-93.6), suggesting 79.5% of the variability in performance is due to differences in study characteristics and only 20.5% due to chance. The random effects model is presented in Figure 2. As reference, the average inter-pathologist agreement from the studies of Bloom et al. [1] (κ = 0.60, 95% CI: 0.53-0.68) and Terry et al. [3] (κ = 0.77, 95% CI: 0.71-0.83) was plotted. Bloom et al. [1] reported inter-observer agreement between 10 pathologists for 126 WSI while Terry et al. [3] provided inter-lab agreement between 17 laboratories for 36 TMA cores. The overall performance was satisfactory; the automated HER2 scoring algorithms agreed with pathologists similar to how well pathologists agreed with each other.

ER Agreement with pathologists: Cohens κ was reviewed for the dichotomized Allred score (Allred $\leq$2: negative, Allred >2: positive). Again, high heterogeneity was present with Higgins I^2=91.0% (95% CI: 68.4-98.4). The summarised performance was κ = 0.74, (95% CI: 0.66-0.83, Figure 3). Here, although the overall agreement of automated scoring systems with pathologists was good, it fell short of the excellent inter-pathologist agreement reported [3] (κ = 0.96, 95% CI: 0.93-0.99).

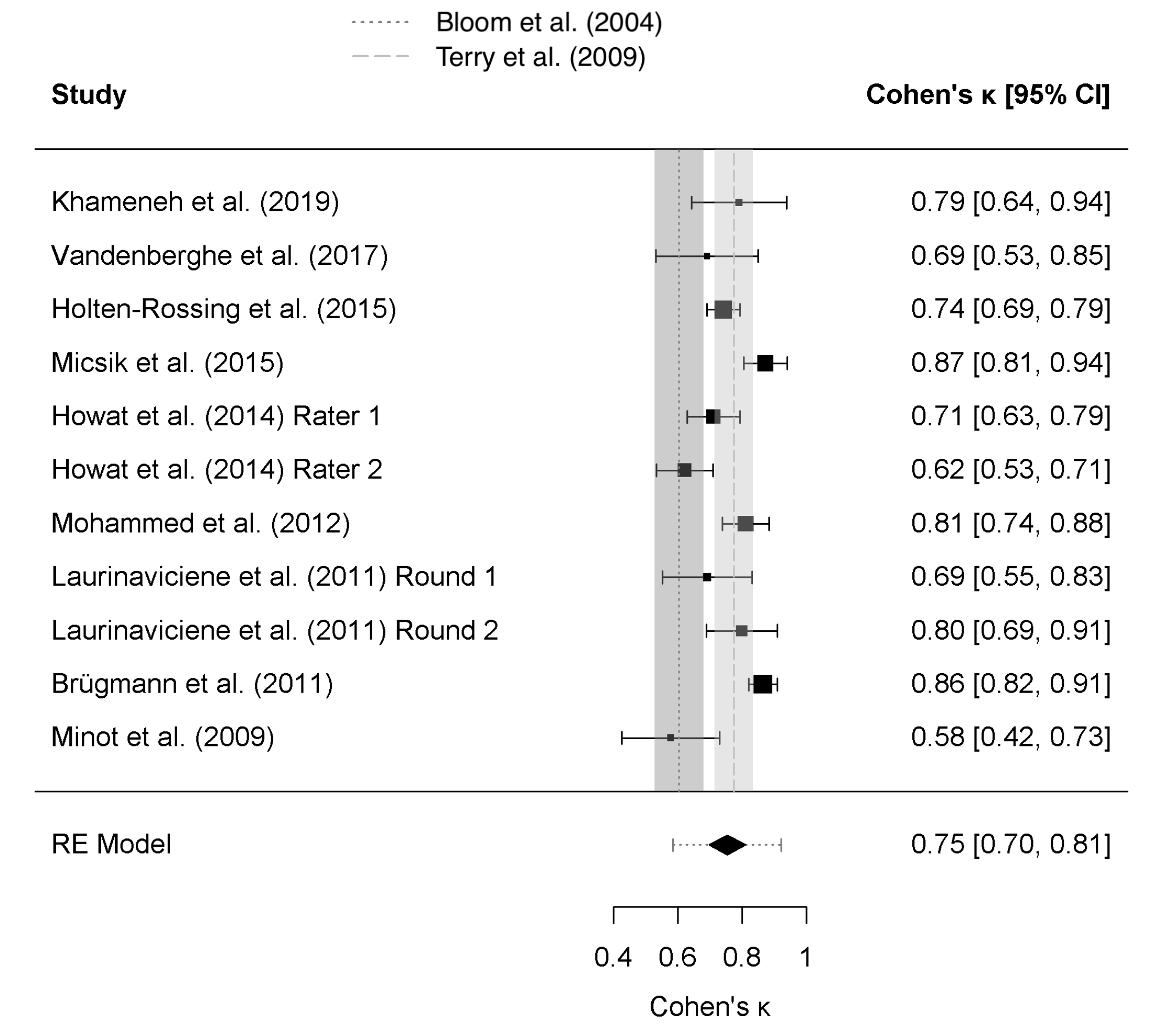

Fig. 2: Random effects meta-analysis of Cohen's κ for HER2 scoring algorithm performance. The size of markers represents the size of the dataset for which performance is reported. All findings cor-respond to a three-class HER2 score (0 or 1+ as negative, 2+ equivocal, 3+ positive).

Results (cont.)

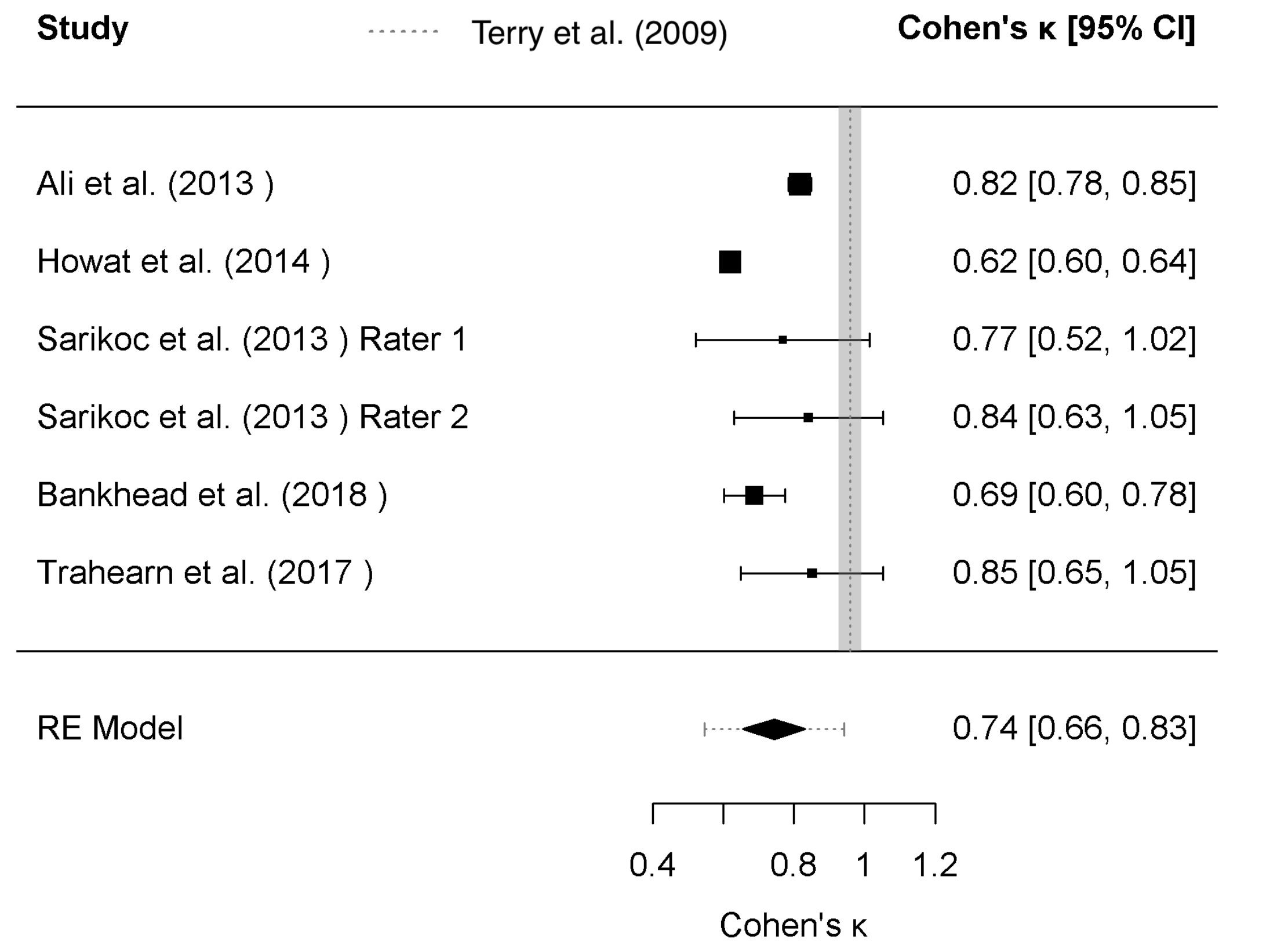

Fig. 3: Random effects meta-analysis of Cohen's κ for ER scoring algorithm performance. The size of markers represents the size of the dataset for which performance is reported. All findings correspond to a dichotomised Allred score ($\leq$2: negative, >2: positive). As benchmark, the human inter-observer agreement for the same task is plotted, using data from Terry et al. [3].

Discussion

Summary of results: Random effects meta-analysis revealed moderately good agreement of HER2 and ER algo-rithms with pathologist scoring (Figures 2, 3). For T-cell markers, a quantitative meta-analysis was impossible, though several studies reported higher correlations between manual and automated scores than between different pathologists.

Limitations: Even though many studies of high quality were retrieved for quantitative meta-analysis, funnel plot analysis revealed a potential publication bias in ER studies. Also, the limited number of studies reporting Cohens κ did not allow subgroup analyses to measure the effect of study characteristics on performance.

Implications: Considerable challenges remain for clinical adoption, with inaccurate tumour area selection accounting for many errors in automated scoring systems and few algorithms available for scoring T-cell markers. Additional challenges are presented in the development of desirable features for routine applications, e.g. interpretability and easy calibration to variable image acquisition settings and staining conditions. Foremost, future validation of automated scoring systems must demonstrate a clear benefit over current standard-of-care by improving reliability, robustness to staining variability or correlation to patient outcome.

References

[1] Kenneth Bloom and Douglas Harrington. *Enhanced Accuracy and Reliability of HER-2/neu Immunohistochemical Scoring Using Digital Microscopy*. 2004. DOI: 10.1309/Y73U8X72B68TMGH5.

[2] Shuyan Sun. "Meta-analysis of Cohen's kappa". In: *Health Services and Outcomes Research Methodology* 11.3 (2011), pp. 145–163.

[3] Jefferson Terry et al. "Implementation of a Canadian External Quality Assurance Program for Breast Cancer Biomarkers: An Initiative of Canadian Quality Control in Immunohistochemistry (cIQc) and Canadian Association of Pathologists (CAP) National Standards Committee/Immunohistochemistry". In: *Applied Immunohistochemistry & Molecular Morphology* 17 (5 2009). ISSN: 1541-2016.