ESC European Society of Cardiology

European Heart Journal (2021) **42**, 2356–2369
doi:10.1093/eurheartj/ehab241

TRANSLATIONAL RESEARCH
Heart failure and cardiomyopathies

An automated computational image analysis pipeline for histological grading of cardiac allograft rejection

Eliot G. Peyster [1*†], Sara Arabyarmohammadi [2†], Andrew Janowczyk[3], Sepideh Azarianpour-Esfahani [3], Miroslav Sekulic [4], Clarissa Cassol [5], Luke Blower[5], Anil Parwani[5], Priti Lal[6], Michael D. Feldman[6], Kenneth B. Margulies [1], and Anant Madabhushi [3]

[1]Cardiovascular Institute, University of Pennsylvania, 3400 Civic Center Blvd, Smilow TRC 11th floor, Philadelphia, PA 19104, USA; [2]Department of Computer and Data Sciences, Case Western Reserve University, 10900 Euclid Avenue, Nord Hall Suite 500, Cleveland, OH 44106, USA; [3]Department of Biomedical Engineering, Case Western Reserve University, 10900 Euclid Avenue, Nord Hall Suite 500, Cleveland, OH 44106, USA; [4]Department of Pathology, University Hospitals Cleveland Medical Center, 11100 Euclid Ave, Cleveland, OH 44106, USA; [5]Department of Pathology, Ohio State University Wexner Medical Center, 450 W 10th Ave, Columbus, OH 43210, USA; [6]Department of Pathology and Laboratory Medicine, University of Pennsylvania, 3400 Spruce Street 6 Founders, Philadelphia, PA 19104, USA

Received 30 November 2020; revised 26 January 2021; editorial decision 8 April 2021; accepted 14 April 2021; online publish-ahead-of-print 13 May 2021

See page 2370 for the editorial comment on this article (doi: 10.1093/eurheartj/ehab226)

Aim

Allograft rejection is a serious concern in heart transplant medicine. Though endomyocardial biopsy with histological grading is the diagnostic standard for rejection, poor inter-pathologist agreement creates significant clinical uncertainty. The aim of this investigation is to demonstrate that cellular rejection grades generated via computational histological analysis are on-par with those provided by expert pathologists

Methods and results

The study cohort consisted of 2472 endomyocardial biopsy slides originating from three major US transplant centres. The 'Computer-Assisted Cardiac Histologic Evaluation (CACHE)-Grader' pipeline was trained using an interpretable, biologically inspired, 'hand-crafted' feature extraction approach. From a menu of 154 quantitative histological features relating the density and orientation of lymphocytes, myocytes, and stroma, a model was developed to reproduce the 4-grade clinical standard for cellular rejection diagnosis. CACHE-grader interpretations were compared with independent pathologists and the 'grade of record', testing for non-inferiority ($\delta = 6\%$). Study pathologists achieved a 60.7% agreement [95% confidence interval (CI): 55.2–66.0%] with the grade of record, and pair-wise agreement among all human graders was 61.5% (95% CI: 57.0–65.8%). The CACHE-Grader met the threshold for non-inferiority, achieving a 65.9% agreement (95% CI: 63.4–68.3%) with the grade of record and a 62.6% agreement (95% CI: 60.3–64.8%) with all human graders. The CACHE-Grader demonstrated nearly identical performance in internal and external validation sets (66.1% vs. 65.8%), resilience to inter-centre variations in tissue processing/digitization, and superior sensitivity for high-grade rejection (74.4% vs. 39.5%, $P < 0.001$).

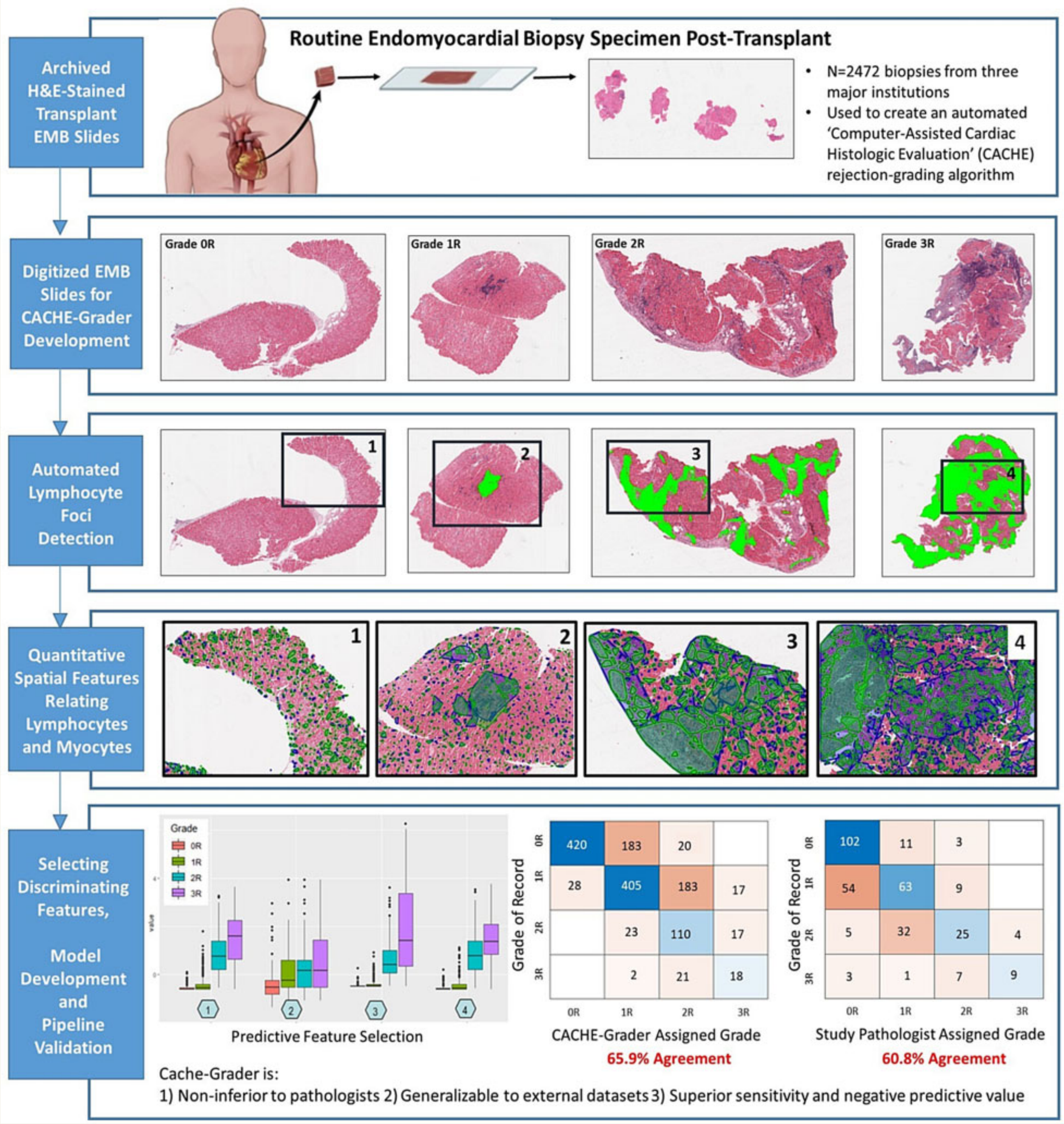
Conclusion

These results show that the CACHE-grader pipeline, derived using intuitive morphological features, can provide expert-quality rejection grading, performing within the range of inter-grader variability seen among human pathologists.

* Corresponding author. Tel: +1 215 554 0993, Email: eliot.peyster@pennmedicine.upenn.edu

† These authors contributed equally to this work.

Graphical Abstract



Archived H&E-Stained Transplant EMB Slides

Routine Endomyocardial Biopsy Specimen Post-Transplant

- N=2472 biopsies from three major institutions
- Used to create an automated 'Computer-Assisted Cardiac Histologic Evaluation' (CACHE) rejection-grading algorithm

Digitized EMB Slides for CACHE-Grader Development

Automated Lymphocyte Foci Detection

Quantitative Spatial Features Relating Lymphocytes and Myocytes

Selecting Discriminating Features, Model Development and Pipeline Validation

Predictive Feature Selection

Grade of Record	0R	1R	2R	3R
0R	420	183	20	
1R	28	405	183	17
2R		23	110	17
3R		2	21	18

CACHE-Grader Assigned Grade

65.9% Agreement

Grade of Record	0R	1R	2R	3R
0R	102	11	3	
1R	54	63	9	
2R	5	32	25	4
3R	3	1	7	9

Study Pathologist Assigned Grade

60.8% Agreement

Cache-Grader is:
1) Non-inferior to pathologists 2) Generalizable to external datasets 3) Superior sensitivity and negative predictive value

Overview of the 'Computer-Assisted Cardiac Histologic Evaluation-Grader' multicentre validation experiment. Nearly 2500 clinical transplant endomyocardial biopsy slides from three transplant centres were used to develop and validate the Computer-Assisted Cardiac Histologic Evaluation-Grader, an automated histological analysis pipeline for assigning standard-of-care cellular rejection grades. The Computer-Assisted Cardiac Histologic Evaluation-Grader performance was compared to both the grade of record and to independent pathologists performing re-grading, demonstrating non-inferiority to expert pathologists, generalizability to external datasets, and excellent sensitivity and negative predictive value.

Keywords Image analysis • Machine learning • Digital pathology • Heart transplant • Allograft rejection

Translational Perspective

This first-in-field, multicentre investigation provides a convincing demonstration of the diagnostic potential of automated histological analysis in cardiovascular and transplant medicine. While automated cellular rejection grading may have direct applications for standardizing histological analyses in multicentre research, future histological analysis systems, which also address antibody-mediated rejection and provide more outcome-based predictions, will be required prior to broad clinical deployment of this emerging methodology. Nevertheless, the degree of accuracy, reliability, interpretability, and generalizability achieved in this multicentre validation experiment is indicative of a maturing diagnostic technology, which is worthy of continued investment and clinical investigation.

Introduction

Heart transplantation is the treatment of choice for end-stage cardiomyopathy that is refractory to medical therapy. Cardiac allograft rejection (CAR) occurs in up to one-third of transplant recipients,[1–3] representing the leading threat to short- and long-term allograft health. As a result, frequent surveillance endomyocardial biopsy (EMB) with histological rejection grading has been included in the International Society for Heart and Lung Transplantation (ISHLT) guidelines since 1990, with recipients typically undergoing 12 or more scheduled EMB procedures in their first year post-transplant alone.[4,5]

Allograft rejection has histopathological features that have been recognized for more than a century.[6] In the modern era, significant effort has been invested into developing standardized metrics for describing these features. To this end, the ISHLT has issued formal CAR histological grading criteria since 1990,[5] aiming to achieve standardization through reductive simplicity. In the case of cell-mediated CAR, the longest-recognized and most prevalent form of rejection, contemporary (ISHLT 2005 guideline) histological grading, utilizes a four-grade scale from '0R' to '3R' based on the number of inflammatory cell 'foci', the extent of cellular infiltration (focal vs. diffuse), and qualitative assessments of lymphocyte 'encroachment' onto myocytes and 'myocyte damage'.[7] Unfortunately, multiple studies have demonstrated poor reliability of ISHLT histological grading for cellular rejection, with a Kappa statistic of 0.39[8] and inter-pathologist agreement of 60–70%, with particularly poor agreement of 28.4% for the higher grades of cellular rejection (2R, 3R), which usually result in major alterations to immunosuppression therapy.[9] The poor statistical performance of the current diagnostic standard has significant clinical and research implications, affecting inter-provider communication, limiting multicentre research, and potentially misguiding therapeutic decisions.

Computational image analysis using 'machine learning' (ML) methodologies can capture and quantify subtle patterns from medical images and has been used to predict disease diagnosis, prognosis, and therapeutic response in oncology, ophthalmology, and dermatology.[10–20] Recently, a computational pathology approach for non-transplant heart tissue achieved excellent diagnostic performance, suggesting the potential value of these methods within cardiovascular diseases.[21,22] In the context of CAR, computational methods provide an opportunity to improve grading consistency and sensitivity via comprehensive and quantitative morphological assessments of EMB specimens.[23] The potential of these methods for transplant EMB analysis has recently been endorsed by the Banff Foundation for Allograft Pathology,[24] though no definitive demonstration of diagnostic performance with translational potential has yet been published.

It is notable that a substantial proportion of the published medical research utilizing ML-enabled image analysis has relied upon 'deep-learning' (DL) approaches using artificial neural networks to examine images.[25] DL methods generate computational models with limited human input during training and provide no clear explanation for their predictions, limiting interpretability. Because of the opacity of this technology,[21,25] DL models are often considered 'black-boxes', a criticism that represents a potential barrier to clinical applications.[26,27] In contrast, models based on 'hand-crafted' feature extraction rely on a foundation of biologically inspired, clearly defined histological principles, and while this approach requires a greater degree of effort and domain-expertise to develop, it may be capable of providing a more transparent and interpretable option than DL methods.[28,29]

In this study, we used a 'hand-crafted' approach to translate the qualitative ISHLT cellular rejection grading criteria into measurable, quantitative variables describing the infiltration of lymphocytes within transplant myocardium. We then employed these variables to build the Computer-Assisted Cardiac Histologic Evaluation (CACHE)-Grader, an automated pipeline for assigning ISHLT cellular rejection grades to digitized EMB histology slides. The goal of this first-in-transplant effort was to demonstrate that lymphocyte-related quantitative features can achieve ISHLT grading performance comparable to that of expert pathologists who represent the field's reference standard. In pursuit of this goal, we sought to validate the CACHE-Grader pipeline within a large, multicentre cohort of EMB slides ($n = 2472$), comparing the CACHE-Grader with the grades of record from the clinical chart and the grades assigned by independent study pathologists (*Graphical Abstract*).

Methods

Study cohort and design

The study cohort was selected from the records at the Hospital of the University of Pennsylvania (UPenn), University Hospitals Cleveland Medical Center (UH), and the Ohio State University Wexner Medical Center (OSU). This cohort consisted of 2472 archived, haematoxylin and eosin (H&E)-stained transplant EMB histology slides, originating from ∼546 transplant recipients. These EMBs were obtained as part of usual care between 2005 and 2018, either as part of scheduled post-transplant surveillance or in a 'for-cause' setting due to suspected rejection, and occurred from 7 to 6415 days after transplant. The EMBs represented the grading efforts of 17 different pathologists of record and encompassed the spectrum of ISHLT cellular rejection grades, including $n = 85$ '3R', $n = 405$ '2R', $n = 979$ '1R', and $n = 889$ '0R' grades. Only EMBs that were deemed 'gradable' and assigned an ISHLT cellular rejection grade were included. By design, the full cohort contained a higher proportion of

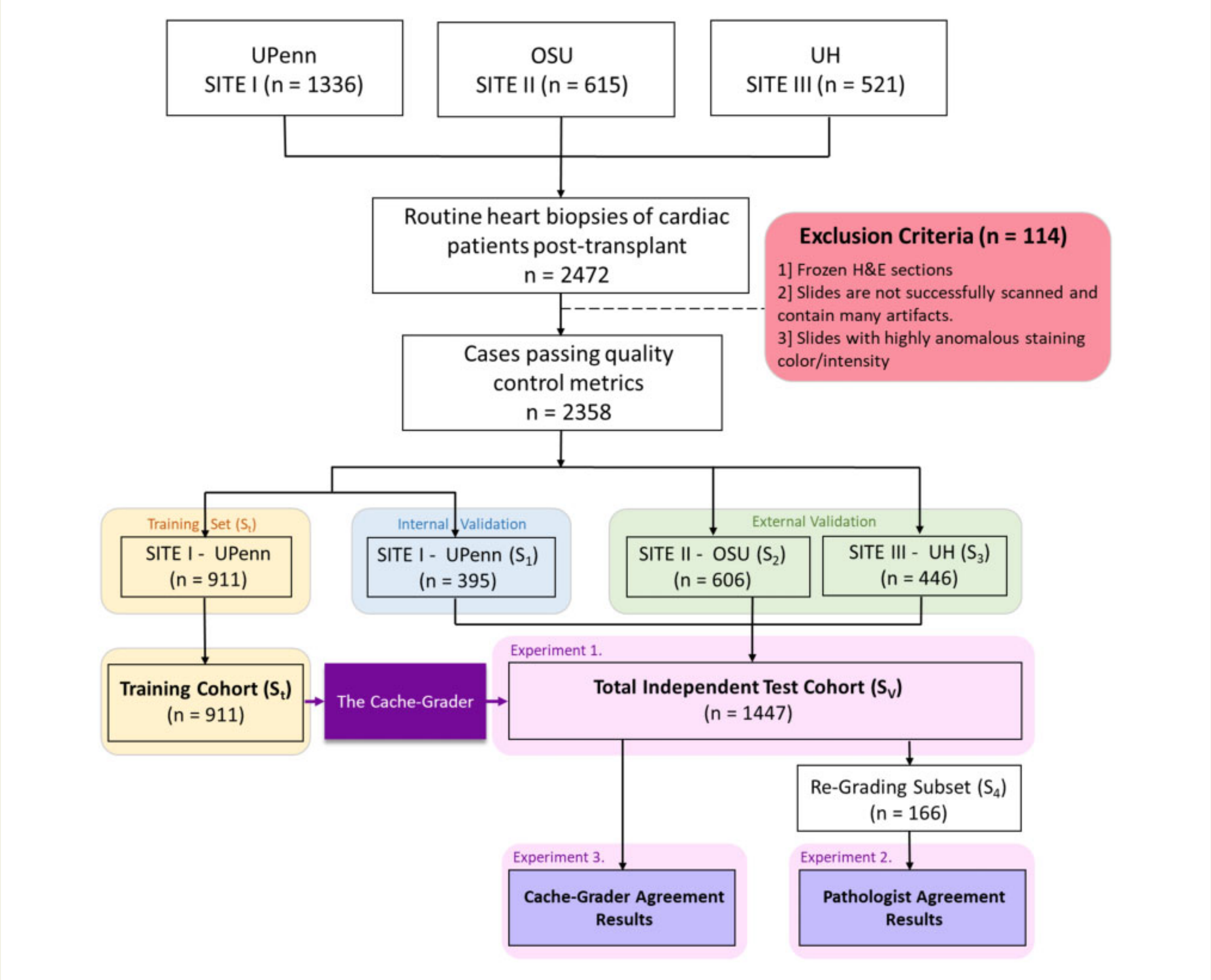

UPenn
SITE I (n = 1336)
OSU
SITE II (n = 615)
UH
SITE III (n = 521)
Routine heart biopsies of cardiac patients post-transplant n = 2472
Exclusion Criteria (n = 114)
1] Frozen H&E sections
2] Slides are not successfully scanned and contain many artifacts.
3] Slides with highly anomalous staining color/intensity
Cases passing quality control metrics n = 2358
Training Set (S_t)
SITE I - UPenn (n = 911)
Internal Validation
SITE I - UPenn (S_1) (n = 395)
External Validation
SITE II - OSU (S_2) (n = 606)
SITE III - UH (S_3) (n = 446)
Training Cohort (S_t) (n = 911)
The Cache-Grader
Experiment 1.
Total Independent Test Cohort (S_V) (n = 1447)
Re-Grading Subset (S_4) (n = 166)
Experiment 3.
Cache-Grader Agreement Results
Experiment 2.
Pathologist Agreement Results

Figure 1 Flowchart outlining the distribution of study histology slides by cohort of origin and experiment.

grades 2R/3R (~20%) than is seen in routine practice (~8%)[9] to ensure adequate power for assessing CACHE-Grader performance in these EMBs of highest clinical importance. For model development and demonstration of predictive performance, the rejection grade assigned to each EMB by the pathologist of record in the medical chart was considered the reference standard. *Figure 1* provides an overview of study EMB slide distribution and utilization, while *Figure 2* provides an overview of the CACHE-Grader pipeline development and deployment. This study was approved by the UPenn institutional review board.

Slide scanning and quality control

All EMBs were collected, processed, and stained per routine workflows at each study centre. Archived H&E-stained slides comprising the study cohort underwent whole-slide scanning at 40× magnification using an Aperio ScanScope or a Philips UltraFast slide scanner (depending on centre). Digitized slides underwent quality control (QC) assessments using HistoQC, an open-source, automated digital pathology analysis software tool for identifying artefacts and measuring slide quality[30] (see Supplementary material online Figures S1 and S2 for details).

Overview of image analysis workflow

Image analysis was conducted on whole-slide EMB images to enable whole-slide grade classification via extraction of histological features inspired by principles of CAR histology.

Tissue compartment segmentation and lymphocyte detection

A stain deconvolution algorithm was applied on tiles of H&E-stained cardiac biopsy images, enabling the determination of areas containing specific stain colours.[31] Following colour deconvolution, a K-means clustering algorithm[32] is used to assign pixels to one of the three tissue-type classes: myocytes, interstitium/stroma, and non-myocyte nuclei (*Figure 3A*).

Isolating lymphocyte clusters

Since lymphocytes relevant to ISHLT grading almost exclusively exist within larger colonies, and because single blue nuclei in isolation can represent several cell types (some leukocytes, some non-leukocytes), identification and isolation of lymphocyte clusters is a necessary process. By

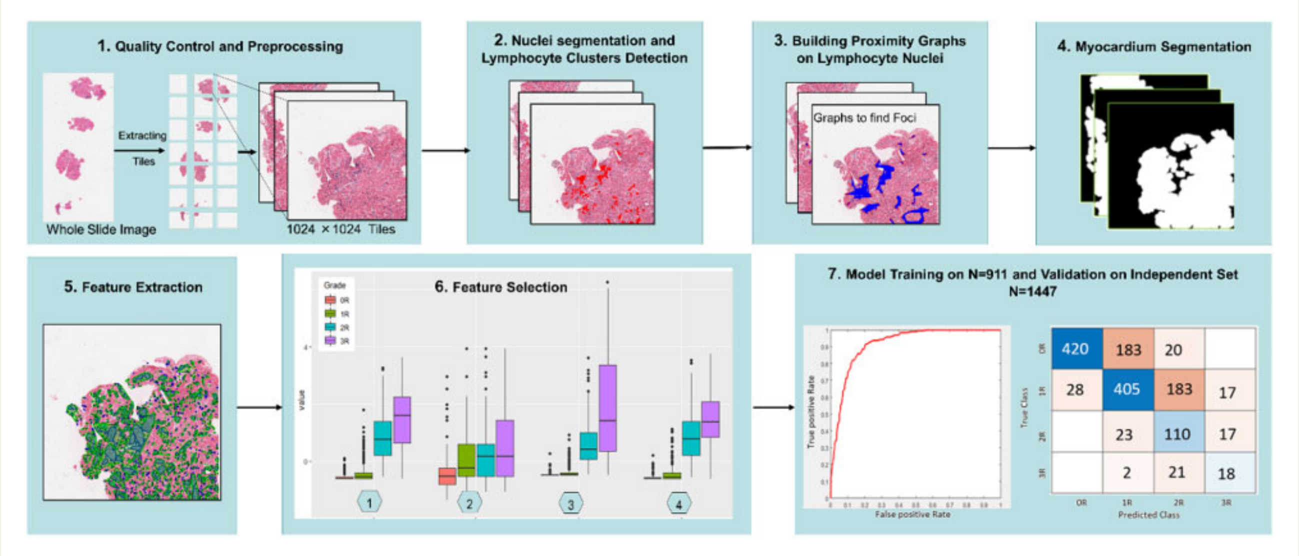

1. Quality Control and Preprocessing
Extracting
Tiles
Whole Slide Image
1024 × 1024 Tiles
2. Nuclei segmentation and Lymphocyte Clusters Detection
3. Building Proximity Graphs on Lymphocyte Nuclei
Graphs to find Foci
4. Myocardium Segmentation
5. Feature Extraction
6. Feature Selection
Grade
0R
1R
2R
3R
Value
1
2
3
4
7. Model Training on N=911 and Validation on Independent Set N=1447
True positive Rate
False positive Rate
True Class
Predicted Class
0R
1R
2R
3R
420
183
20
28
405
183
17
23
110
17
2
21
18

Figure 2 Summary of Computer-Assisted Cardiac Histologic Evaluation-Grader pipeline development, including image analysis, feature selection, model training, and grading performance validation. The Computer-Assisted Cardiac Histologic Evaluation-Grader pipeline involves computerized identification and multi-parameter quantitation of immune cell infiltrates, measuring size, counts, densities, and spatial relationships (panels 1–4, see *Figure 3* for more detailed image feature extraction workflow examples). Following feature extraction (panel 5), features were assessed and ranked based on ability to discriminate between International Society for Heart and Lung Transplantation histological grades (panel 6). In the descending order of discriminatory ability, features were added to the Computer-Assisted Cardiac Histologic Evaluation-Grader predictive models until optimal performance was achieved on the training sets. These features were subsequently 'locked down' into the final Computer-Assisted Cardiac Histologic Evaluation-Grader models which were deployed in the study validation set (panel 7).

performing disc-dilation and area-thresholding, lymphocyte clusters/colonies meeting relevant size criteria are identified while isolated and potentially misleading single-nuclei are omitted (*Figure 3B*).

From clusters to foci

Detecting lymphocyte foci is essential to ISHLT grading, forming the basis of grade differentiation. For each lymphocyte cluster identified, proximity graphs are built on lymphocytes within *and* between clusters, facilitating the reproducible 'lumping' vs. 'splitting' of adjacent clusters into foci via thresholding of Euclidean distances (*Figure 3B*). The threshold parameters of the proximity graphs were initially set based on manually labelled examples of foci and then refined in an automated fashion based on a subset of graded training-set slides.

Identifying foci neighbourhoods

The local neighbourhood of a lymphocyte focus informs histological grading, with lymphocytes (i) within endocardium, (ii) neatly within interstitium, and (iii) encroaching upon myocyte borders being explicitly mentioned in the ISHLT criteria. Myocyte and interstitium segmentations were achieved via K-means clustering (as described above), while more gross discrimination between myocardium vs. endocardium compartments was achieved via a disc-dilation method (*Figure 3A*, right panel). Spatial analysis of the locations and edge-interactions of lymphocyte clusters/foci were then performed, as illustrated in *Figure 4*.

Data analysis and statistical methods

Feature selection and classifier construction

Model development and iterative calibration of the CACHE-Grader pipeline was conducted in a training subset (S_t) of 911 randomly selected slides from Site I (UPenn). Through execution of the study image analysis workflow, a set of 154 domain-inspired, hand-crafted, quantitative histological features were extracted from each study image. Broadly, these features pertained to three categories: (i) features quantifying number of lymphocyte clusters/foci in different tissue compartments, (ii) size/density statistics for lymphocyte clusters, both raw and normalized by size of relevant tissue compartments, and (iii) the spatial/edge interactions of lymphocyte clusters and foci (e.g. foci edges encroaching upon myocyte borders, foci constrained to the interstitium). Specific features are described in Supplementary material online, *Table S1*, along with a tabular summary of conventional ISHLT cellular rejection grading criteria for reference (Supplementary material online, *Table S2*). The extracted features were used to develop two predictive models which would generate the final CACHE-Grader outputs: one for providing binary low-grade (0R, 1R) vs. high-grade (2R, 3R) classification (M_1), and the other for providing clinical standard 4-grade classification (M_2). Several different linear classifiers were employed and compared for generating predictive models, with a support vector machine (SVM) classification method ultimately selected for developing both the M_1 and M_2 final models (Supplementary material online, *Table S3*). For creating M_1, a Wilcoxon rank-sum test was used to select top predictive features, with optimal model performance achieved with an SVM using only two features. For M_2, a *t*-test was employed to rank predictive features, with an SVM using 15 features achieving optimal predictive performance while minimizing over-fitting.[33]

Experiment 1: internal and external validation

CACHE-Grader validation was performed on slides from all three study sites, using the image analysis workflow and predictive models developed during training. It should be noted that no further alterations to the image analysis pipeline or predictive models occurred upon completion of CACHE-Grader training and the commencement of validation

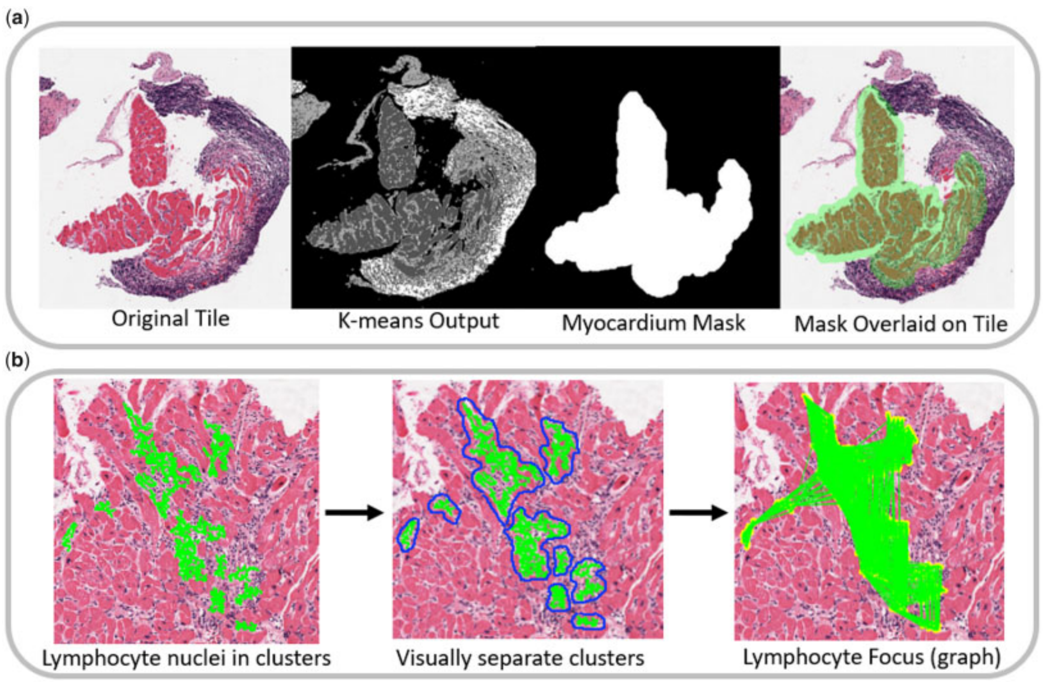


Figure 3 Computer-Assisted Cardiac Histologic Evaluation-Grader feature extraction approach. (*A*) Workflow for compartment segmentation. Left panel: a digitized clinical histology slide from EMB tissue stained in haematoxylin and eosin. Middle-left: K-means segmentation into myocytes (dark grey), interstitium/stroma (light grey), and non-myocyte nuclei (white). Middle-right: Dilated myocyte mask created from myocyte segmentation, identifying the myocardial compartment. Right: Overlay of myocardial mask onto original tile, demonstrating the ability to independently analyse lymphocytes within the myocardial vs. endocardial compartments (which in this example contains a Quilty lesion rather than an infiltrating lymphocyte focus). (*B*) Workflow for lymphocyte foci identification. Left panel: green = lymphocytes identified as 'clustering' together via area-thresholding of individual lymphocyte nuclei (overlay of individual lymphocytes in green which comprise a cluster). Middle panel: blue outline = Distinct lymphocyte clusters identified for feature extraction. Right panel: Applying proximity graph thresholding to lymphocyte clusters allows merging of nearby clusters into a common 'lymphocyte focus' for reproducing the foci counting as outlined in the International Society for Heart and Lung Transplantation histological rejection grading scheme.

experiments. The remaining 395 slides from Site I not used during CACHE-Grader training represented an internal validation set (S_1), while the 1052 slides from sites II (S_2) and III (S_3) represented two external validation sets. The external validation sets enabled assessments of CACHE-Grader generalizability, both to groups of pathologists beyond those who provided grades of record in the training set (as these same pathologists contributed grades of record in the internal validation set), and to alternative pathology lab workflows for slide staining and scanning that may create variability in slide appearance across centres. In total, 1447 slides were available as a composite validation set (S_V) (*Figure 1*). Performance of the CACHE-Grader was assessed based on the percent agreement of M_1 and M_2 outputs with the grade of record. Separate analyses were performed for internal, external, and combined validation sets. Additional measurements including Cohen's kappa (raw and quadratic weighted)[34] and, for binary model M_1, area under the receiver operating characteristic curve (AUC), sensitivity, and specificity were also calculated. Uniform manifold approximation and projection (UMAP) embedding was also employed to further assess the resilience of CACHE-Grader pipeline to pre-analytical variations in slide appearance/quality.

Experiment 2: inter-pathologist agreement

A randomly selected subset ($n = 166$) of S_V slides, referred to as the S_4 set, were re-graded by three independent study pathologists to assess inter-pathologist agreement within the study cohort. Inter-grader percent agreement, along with Cohen's kappa (raw and quadratic-weighted), was calculated in comparison to the grade of record—the same reference standard used to test the grading performance of the CACHE-Grader in Experiment 1.

Experiment 3: non-inferiority determination

Due to well-documented inter-grader variability, true diagnostic 'accuracy' for rejection grading is difficult to define and difficult to test experimentally.[8,9] Recognizing that conventional accuracy and superiority testing are challenging to interpret in the absence of a high-performing, objective 'gold standard', we performed comparisons of the CACHE-Grader performance from experiment 1 to the performance of study pathologists in experiment 2, testing for non-inferiority. Fundamentally, the purpose of this experiment was to demonstrate that the CACHE-Grader performs within the expected and established ranges of inter-grader variability seen among human experts in real-world practice, while also demonstrating strong reliability and generalizability across a spectrum of providers, centres, and eras.

In the large CARGO-II study, composite inter-pathologist percent agreement on nearly 500 EMB samples undergoing ISHLT cellular rejection grading was 70.7% [95% confidence interval (CI) 67.7–73.7%], with most paired-agreements falling within 4–6% of one another.[9] Based on

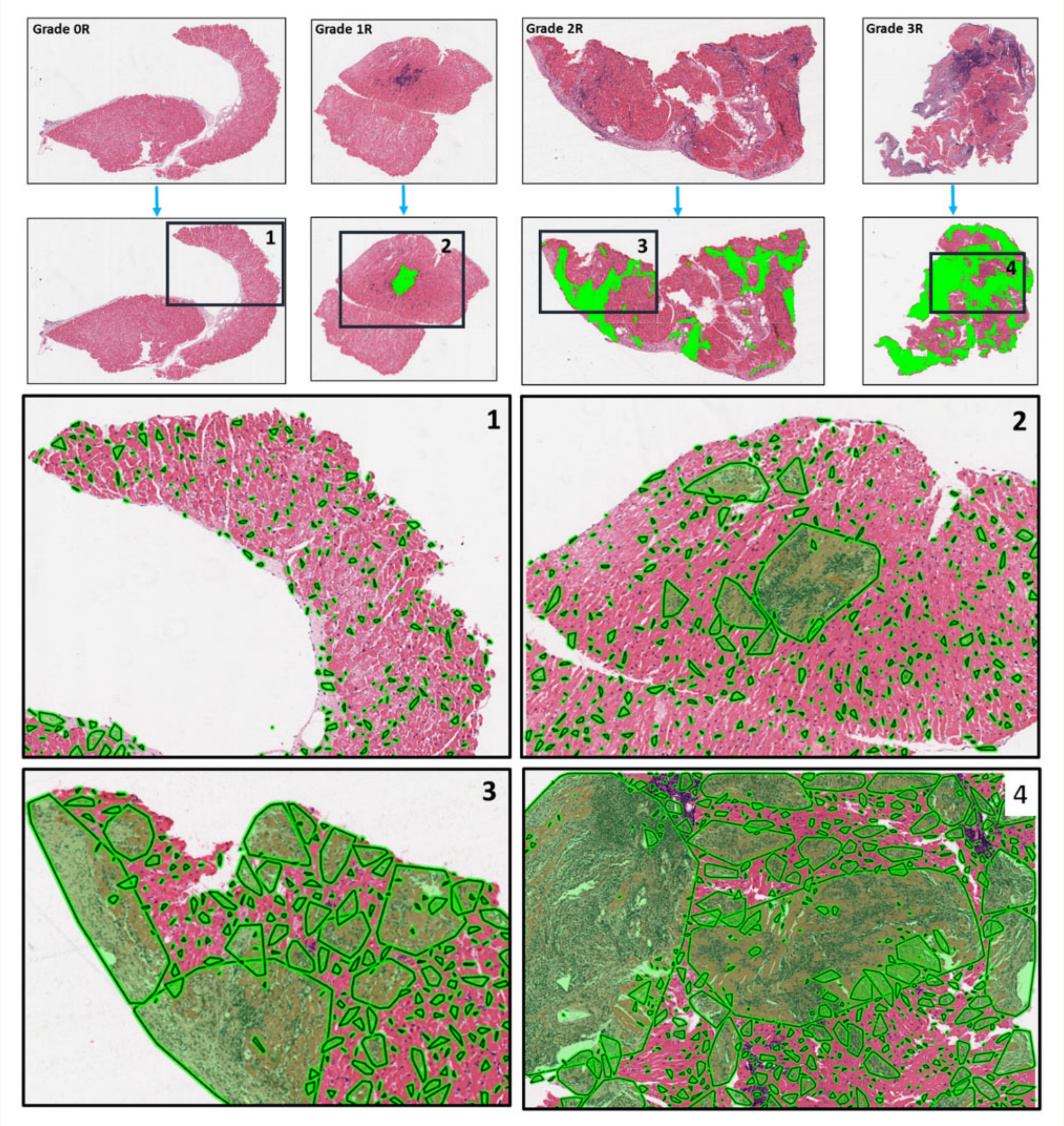


Figure 4 By-grade examples of image analysis results: The first row shows biopsy specimen of different rejection grades. The second row demonstrates the proximity graphs (i.e. foci) in green built across the tissue over lymphocytes. The third and fourth rows demonstrate that the clusters of proximally situated lymphocytes are different in shape and size between high and low rejection grades and also between different grades. One may appreciate that in 2R and 3R cases (high-grade rejection cases), the lymphocyte clusters covered most of the tissue while in 0R and 1R (low-grade rejection cases), lymphocyte clusters are dispersed, small, and they cover only a small proportion of the tissue specimen.

this large historical result, we pre-specified a significant margin of difference for excluding non-inferiority to be an absolute pair-wise agreement difference >6%.

Comparisons of additional agreement statistics, including Cohen's kappa and intraclass correlation coefficients[35] for all pair-wise combinations of graders were also conducted to supplement non-inferiority testing. These included assessments of paired-agreement between the CACHE-grader and individual study pathologists, and between pairs of study pathologists themselves when commonly-graded slides were available. All statistical analyses were conducted in Stata IC v.15.0 (StataCorp LLC).

Table 1 Summary of Computer-Assisted Cardiac Histologic Evaluation-Grader performance in study validation sets

Validation set	Total slides (*n*)	Binary classification (M_1)		4-Grade classification (M_2)	
		Correctly assigned (*n*)	% Agreement (95% CI)	Correctly assigned (*n*)	% Agreement (95% CI)
Internal (S_1)	395	350	0.886 (0.851–0.916)	261	0.661 (0.612–0.707)
External ($S_2 + S_3$)	1052	872	0.829 (0. 805–0.851)	692	0.658 (0.628–0.686)
Combined (S_V)	1447	1222	0.845 (0.825–0.863)	953	0.659 (0.634–0.683)

CI, confidence interval.

Table 2 Computer-Assisted Cardiac Histologic Evaluation-Grader M_2 performance, by grade

ISHLT grade	Total slides (*n*)	Correctly assigned (*n*)	% agreement (95% CI)
0R	623	420	0.674 (0.636–0.711)
1R	633	405	0.640 (0.601–0.677)
2R	150	110	0.733 (0.655–0.802)
3R	41	18	0.439 (0.285–0.603)

CI, confidence interval; ISHLT, International Society for Heart and Lung Transplantation.

Results

Experiment 1: internal and external validation of the Computer-Assisted Cardiac Histologic Evaluation-Grader

In the composite validation set (S_V), the CACHE-Grader M_1 for providing high- vs. low-grade classification achieved a percent agreement with the grade of record of 84.5%, with an AUC of 0.92, a sensitivity of 0.85, and a specificity of 0.84. The performance of M_1 within internal validation set S_1 was similar to, though better than, the agreement in external sets $S_2 + S_3$ (88.6% vs. 82.9%, $P = 0.008$), suggesting reasonable generalizability of the model to external data. With regard to model M_2 for four-grade classification, the CACHE-Grader achieved a percent agreement of 65.9% within S_V, with very similar performance in set S_1 vs. sets $S_2 + S_3$ (66.1% vs. 65.8%, $P = 0.91$), further supporting generalizability. There was no significant difference ($P = 0.22$) in CACHE-Grader performance by era, with a 63.0% (208/330) agreement for older EMBs (years 2005-2011) vs. 66.7% (745/1117) for newer EMBs (years 2012-2018). The complete results for CACHE-Grader performance are summarized in *Tables 1 and 2* and *Figure 5*.

Experiment 2: inter-pathologist agreement

To further validate the CACHE-Grader, a representative subset of validation set slides (S_4) underwent blinded re-grading by three independent study pathologists. In total, study pathologists provided 328 independent re-grades of slides in S_4. Inter-pathologist percent agreement results via comparison to the grade of record, along with Cohen's kappa statistics, are summarized in *Table 3*, with companion confusion matrices in Supplementary material online, *Figure S3*. Overall, the pathologist agreement with the grade of record was fair, with an averaged percent agreement of 60.7% (95% CI 55.2–66.1%), a composite raw kappa of 0.41 (95% CI 0.34–0.48), and a composite quadratic-weighted kappa of 0.65 (95% CI 0.55–0.75). Consistent with prior reports, percent agreement was far better for grade 0R (87.9%, 95% CI 80.7–93.3%) than for other grades (*Table 3*), with particularly poor cumulative pair-wise agreement for grades 2R and 3R. When results from all possible pairs of pathologists are tabulated ($n = 495$ total available grading pairs, inclusive of pairs of study pathologists performing re-grading on the same digital slides), the composite inter-pathologist percent agreement remains similar at 61.5% (95% CI 57.0–65.8%) with no significant changes in kappa statistics (*Table 4*).

Experiment 3: comparison of Computer-Assisted Cardiac Histologic Evaluation-Grader and expert pathologists

Comparisons of the CACHE-Grader agreement with the grade of record from the composite validation set S_V (*Table 1*) to the inter-pathologist agreement results from the re-grading set S_4 (*Table 3*) support a conclusion of non-inferiority in this experiment. The lower bound of the 95% CI for overall CACHE-Grader percent agreement is easily within the pre-specified 6% margin of difference when compared to the combined study pathologist percent agreement. In fact, the 95% CI for the CACHE-Grader suggests superiority given it does not include the combined study pathologist percent agreement value ($P = 0.002$). When all possible pathologist-to-pathologist comparisons and all possible CACHE-Grader-to-pathologist combinations are considered as in *Table 4* (including not just paired comparisons with grade of record, but also study pathologist/study pathologist pairs and CACHE-Grader/study pathologist pairs), the CACHE-Grader remains non-inferior, but now without any suggestion of superiority (61.5% vs. 62.6%, $P = 0.66$).

When the *highest* grade assigned to a slide in S_4 by *any* pathologist (pathologist of record or study pathologist) is compared to the CACHE-Grader, the percent agreement is 69.1% (95% CI 62.5–76.6%), significantly higher than the overall inter-pathologist agreement ($P = 0.046$). This can be considered evidence of the high sensitivity of the CACHE-Grader for serious histological rejection, a finding supported by the nearly twice as high percent agreement of

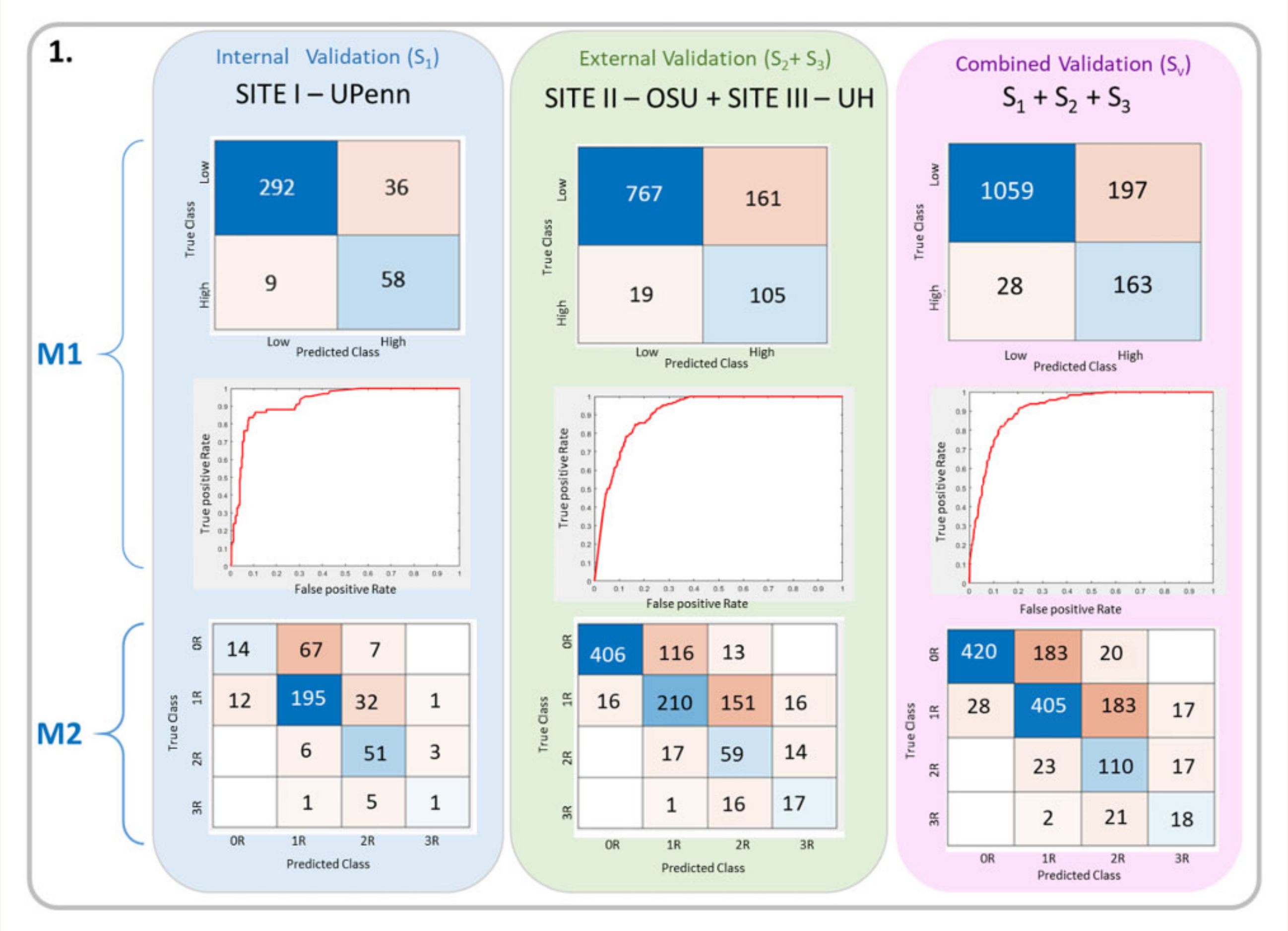


Figure 5 Computer-Assisted Cardiac Histologic Evaluation-Grader validation results: confusion matrices and Receiver-operating characteristic curves demonstrating M_1 and M_2 performance on independent validation sets (S_1, S_2, and S_3).

Table 3 Performance of study pathologists and the Computer-Assisted Cardiac Histologic Evaluation-Grader within S_4

	Pathologist 1	Pathologist 2	Pathologist 3	Combined	CACHE-Grader
Grade 0R, % agreement (*n*)	0.929 (39/42)	0.743 (26/35)	0.949 (37/39)	0.879 (102/116)	0.746 (44/59)
Grade 1R, % agreement (*n*)	0.564 (22/39)	0.605 (23/38)	0.367 (18/49)	0.500 (63/126)	0.762 (48/63)
Grade 2R, % agreement (*n*)	0.529 (9/17)	0.238 (5/21)	0.393 (11/28)	0.379 (25/66)	0.788 (26/33)
Grade 3R, % agreement (*n*)	0.400 (2/5)	0.667 (4/6)	0.500 (2/4)	0.45 (9/20)	0.600 (6/10)
All grades, % agreement (95% CI) (*n*)	0.699 (0.601–0.786) (*n* = 72/103)	0.580 (0.479–0.678) (*n* = 58/100)	0.552 (0.461–0.641) (*n* = 69/125)	0.607 (0.552–0.661) (*n* = 199/328)	0.752 (0.678–0.815) (*n* = 124/165)
Cohen's kappa	0.54 (0.40–0.67)	0.38 (0.25–0.51)	0.35 (0.24–0.45)	0.41 (0.34–0.48)	0.64 (0.54–0.74)
Quadratic kappa	0.72 (0.54–00.92)	0.65 (0.46–0.85)	0.60 (0.45–0.75)	0.65 (0.55–0.75)	0.85 (0.70–0.99)

CACHE, Computer-Assisted Cardiac Histologic Evaluation; CI, confidence interval.

the CACHE-Grader with the grade of record for 2R/3R rejection when compared to study pathologists within S_4 (39.5% vs. 74.4%, $P < 0.001$). Related to this higher sensitivity, the CACHE-Grader also demonstrates a much lower 'false-negative' rate, *never* assigning a 0R to an EMB with a grade of record of 2R or 3R (0/201), compared to a 9% rate (8/86) for study pathologists.

Table 4 Comparison of study pathologist agreement statistics and Computer-Assisted Cardiac Histologic Evaluation-Grader agreement

	Inter-pathologist agreement	CACHE-pathologist agreement
Combined % agreement (95% CI) (*n*)	0.615 (0.570–0.658) (*n* = 302/491)	0.626 (0.603–0.648) (*n* = 1111/1775)
Average Cohen's kappa (95% CI)	0.42 (0.39–0.45)	0.44 (0.41–0.47)
Average quadratic kappa (95% CI)	0.67 (0.57–0.77)	0.65 (0.61–0.70)
Intraclass correlation coefficient	0.66 (0.59–0.74)	0.65 (0.62–0.70)

CACHE, Computer-Assisted Cardiac Histologic Evaluation; CI, confidence interval.

Discussion

In this work, we presented a novel computational approach to histological rejection grading of transplant EMB tissues. Starting with historical evidence of poor inter-grader agreement using ISHLT criteria and an unmet need for more reproducible grading approaches, we successfully demonstrated that an automated image analysis pipeline can provide cellular rejection grading with statistical performance on-par with that of expert pathologists.

Comparisons with other computational histology research

The present research represents by far the largest application of computational histological analysis within cardiovascular medicine or solid organ transplant medicine to date. In contrast, within the tissue-rich field of oncology, ML approaches for digital pathology are more established, presenting better opportunities for comparison to this experiment.[20,36–40] Two investigations using DL methods to perform histological grading of prostate biopsies[39,40] were recently published, with multicentre validation results showing performance on-par with expert pathologists much as we have shown in the present work. Perhaps the closest research related to the work presented in this study is that of Nirschl *et al.*[22] where an ML approach was used to analyse heart tissue samples for the presence or absence of clinical heart failure. However, all the aforementioned and cited studies have relied upon DL methodologies employing neural network models. Neural networks use largely unsupervised feature generation approaches, performing automated and repeated image transformations to find representations that best distinguish categories of interest. While this approach is powerful and does not require subject-specific expertise to employ, the abstract and automated nature of DL image transformations often preclude any clear, biologically meaningful explanation of the image features that are responsible for model predictions.

Strengths of our approach

Recognizing that adoption of automated image analysis platforms into clinical practice will require not only convincing statistical demonstrations but also clear, transparent, and biologically inspired methods, we chose to employ a 'hand-crafted' feature-engineering approach in this work.[28] This approach relied on careful supervision during the initial feature extraction process, which generated feature maps based on fundamental principles of rejection histopathology. The features employed in the CACHE-Grader are derived from the interactions between myocytes and lymphocytes, describing foci counts, lymphocytic infiltrate distribution/extent, and various cellular/compartment interactions in a quantitative manner. The extraction of these domain-relevant histological features allows the CACHE-Grader to assign rejection grades in a manner that has clear parallels with human grading workflows. Perhaps as a result of the strong biological foundation, the final CACHE-Grader required only a relatively small number of quantitative lymphocyte and myocyte features to achieve optimal ISHLT grading performance. The relative simplicity of these models stands in contrast to DL approaches, which tend to produce models that are much less parsimonious, possibly because there is no initial biological foundation which helps prevent the inclusion of noisy, artefactual, or nonsensical variables.

Interpretation of findings

The CACHE-Grader achieved similar percent agreements within both the internal and external validation sets in this experiment, suggesting that the models are both generalizable to external data and robust, capable of strong performance even when different slide scanning and tissue processing workflows are applied. The UMAP plots showing unbiased two-dimensional representations of all image quality metrics and all quantitative rejection features provide further support for the resilience of the CACHE-Grader to site-specific batch effects (Supplementary material online, *Results* and Supplementary material online, *Figures S4* and *S5*). Comparing the combined CACHE-Grader-to-pathologist agreement to the combined inter-pathologist agreement clearly supports a conclusion of non-inferiority, with nearly identical agreement metrics. Interestingly, within the re-grading S_4 subset where there is full overlap of graded slides between the CACHE-Grader and the study pathologists, the CACHE-Grader actually has a significantly higher percent agreement with the grade of record than do the independent pathologists. This is likely due to the by-chance higher percentage of high-grade EMBs in S_4 as compared to the total cohort (26% vs. 14%), along with the CACHE-Grader's demonstrated high sensitivity for high-grade rejection.

The top-ranked features comprising the CACHE-Grader describe the counts of and (normalized) areas covered by lymphocyte foci found within the myocardial compartment. This is unsurprising, as estimates of foci count and infiltrate extent represent core criteria of the ISHLT cellular rejection grading framework.[7] The specific importance of lymphocyte activity within the myocardial compartment also suggests a relative lack of predictive importance for lymphocyte activity outside the myocardium. 'Extra-myocardial' lymphocyte activity largely consists of endocardial infiltrates corresponding to 'Quilty' lesions, and lymphocyte activity found within areas of prominent vasculature, corresponding to 'peri-vascular' infiltrates.[7,8,41] Since these histological processes play a minor part within the ISHLT criteria for discriminating between different histological grades of cellular rejection,[7] it is understandable that quantitative features associated with these processes were not valued highly by the models comprising the CACHE-Grader. Finally, it is notable that quantitative features describing lymphocyte activity 'respecting' vs. spatially 'encroaching' upon myocyte borders were not required to achieve strong predictive models. The concept of myocyte encroachment/damage, while codified within the ISHLT grading criteria and of intuitive mechanistic importance, has also long been recognized as a particularly subtle, vague, and subjective finding.[42,43] Given this, it is conceivable that the degree of variability in pathologist assessments of encroachment/damage is substantial enough to render associated quantitative features unhelpful when designing models to reproduce pathologist quality ISHLT grades. This does not mean that such features are without biological value, but rather that they do not add value for this particular modelling task. Understanding the true value of these myocyte/lymphocyte relational features may require models designed to predict more clinical or outcome-based endpoints.

In the S_4 re-grading set, it is noteworthy that there were several episodes of 'major discordance' in which one human grader assigned a grade that was at least two grade-points different from another grader. This occurred in 3.0% (15/491) of paired human-grader comparisons, affecting 6% (10/165) total slides in the S_4 set. While it is initially surprising to see such large differences in grading assessments, this finding has been seen in prior published work. In the largest study on inter-grader agreement using data from the CARGO-II clinical trial,[9] 11/59 (18.6%) of 'high-grade' EMB slides as determined by the pathologist of record were re-graded 0R by a centralized panel of trial pathologists. The most likely primary cause of these major discordance events pertains to the determination of infiltrating 'Quilty' lesions. These infiltrates of endocardial origin can be quite large, can invade into the myocardium, and depending on interpretation, can vastly affect grade results. It is notable that 9/10 slides with major discordance in S_4 had Quilty lesions noted by at least one pathologist. Example slides resulting in major discordance are shown in Supplementary material online, *Figure S6*.

Limitations

There are several notable limitations of the present work. In clinical workflows, pathologists examine slides under microscopes and, in challenging cases, may choose to examine serial sections before arriving at a final grade. It is possible that study pathologists, who are not accustomed to digital pathology and 'one slide, one grade' workflows, may have had their performance hindered by these departures from conventional practice. This might explain why the CACHE-Grader tended to have slightly better performance than the pathologists throughout the various study analyses. Arguing against this is the fact that study pathologist agreement when compared to one another (grading via digital, 'one slide, one grade' workflows) vs. when compared to the grade of record (graded via conventional workflows) was identical at 60.8%.

Another notable, and unavoidable, limitation of this study is reliance on the cellular rejection grade of record as the 'diagnostic truth' label for CACHE-Grader training. Although ISHLT histological grading is the accepted diagnostic standard, it is an imperfect standard and does not represent objective diagnostic truth. As we have demonstrated in this study, there is significant inter-grader variability in this field. Given that the grades of record used for both training and validation represent the work of many different individual pathologists, it is certain that different pathologists providing grades of record would have assigned different grades to specific EMBs in this experiment. As a result, there is no single, consistent grading 'ruleset' for the CACHE-Grader to learn from during training, nor an irrefutable standard to compare predictions to during validation. This imposes an upper-limit on the degree of percent agreement that is achievable for the CACHE-Grader, as it will always be constrained by the degree of inter-pathologist agreement within the image set used to train it. This limitation is inherent to not only the question of histological rejection but also many similar diagnostic scenarios in pathology and medical imaging for which the reference standard is derived from expert visual interpretation.[28,44,45] It is also the justification for pursuing non-inferiority study designs, which permit an assessment of whether performance is comparable to that of a typical expert evaluator.

Finally, it should also be noted that the CACHE-Grader was only developed to provide assessments of cellular rejection. Antibody-mediated rejection (AMR) is an important clinical entity as well, but substantial inter-centre variation in the frequency and approach to AMR screening, along with the need for immune-staining and serologic testing for proper evaluation,[46] make AMR challenging to address in this multicenre cohort. While computational image analysis methods should be similarly well-suited to the task of AMR diagnosis as they are to cellular rejection grading, a carefully designed cohort with a harmonized AMR diagnostic protocol would likely be required to perform an appropriate validation experiment.

Translational implications

In light of the CACHE-Grader's combination of consistency and potential for remote (cloud-based) accessibility, the CACHE-Grader could function as a 'core lab' for standardized grading in future multicentre research. It is also tempting to envision clinical translation of the CACHE-Grader, though it is quite unlikely to replace the role of pathologists in rejection grading. In complex cases, expert pathologists provide additional annotations and perspectives beyond cellular rejection grade alone (e.g. AMR-related features, as discussed above), something the CACHE-Grader is not currently designed to do. The CACHE-Grader may also struggle with slides containing major processing/staining artefacts, at least until automated QC is more seamlessly connected to the pipeline. Slides with major processing/staining artefacts, which are deemed to abnormal by automated QC, would also necessarily require a trained pathologist to evaluate. Instead or replacing the pathologist, given the high sensitivity for high-grade rejection and low false-negative rate, the CACHE-Grader may someday be deployed as a screening tool or quick-reference 'second opinion' in clinical practice. Finally, it is worth noting that the quantitative histological rejection features generated by the image analysis workflow may also be useful for predicting more granular patient-level outcomes such as overt allograft injury or future rejection risk. Such applications could prove clinically valuable and could help move the field beyond conventional grading, establishing histological criteria, which are anchored in more clinically meaningful metrics.

Conclusion

The CACHE-Grader is the first rigorously validated tool for automated histological diagnosis in cardiovascular or solid organ transplant medicine. In this multicentre validation experiment, the CACHE-Grader demonstrates statistical non-inferiority to the field's diagnostic standard, excellent sensitivity, generalizability to external datasets, and resilience to variations in slide processing. This convincing statistical performance, along with the clear, biological principles underlying it, should inspire confidence in future end-users of the CACHE-Grader and future pipelines built using these methods.

Supplementary material

Supplementary material is available at *European Heart Journal* online.

Funding

Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under award numbers 1U24CA199374-01, R01CA202752-01A1, R01CA208236-01A1, R01 CA216579-01A1, R01 CA220581-01A1, 1U01 CA239055-01, 1U01CA248226-01, and P30 CA16058 and by the National Institute for Biomedical Imaging and Bioengineering 1R43EB028736-01, the National Center for Advancing Translational Sciences of the National Institutes of Health under awards number TL1TR001880 and KL2TR001879, the National Heart, Lung and Blood Institute under award number R01HL151277-01A1, and the Gund Family Fund at the University of Pennsylvania. Additional support was provided by the National Center for Research Resources under award number 1 C06 RR12463-01, VA Merit Review Award IBX004121A from the United States Department of Veterans Affairs Biomedical Laboratory Research and Development Service, The Department of Defense Breast Cancer Research Program Breakthrough Level 1 Award W81XWH-19-1-0668, The Department of Defense Prostate Cancer Idea Development Award (W81XWH-15-1-0558), The Department of Defense Lung Cancer Investigator-Initiated Translational Research Award (W81XWH-18-1-0440), The Department of Defense Peer Reviewed Cancer Research Program (W81XWH-16-1-0329), The Ohio Third Frontier Technology Validation Fund, The Wallace H. Coulter Foundation Program in the Department of Biomedical Engineering, the Clinical and Translational Science Award Program (CTSA) at Case Western Reserve University, and The Ohio State University Comprehensive Cancer Center Comparative Pathology & Digital Imaging Shared Resource. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, the U.S. Department of Veterans Affairs, the Department of Defense, or the United States Government.

Contributors

E.G.P. conceived the study along with A.M. and K.B.M. S.A., A.J., and S.A.-E. conducted image analysis, with supervision, method support, and result validation from E.G.P., A.M., A.J., M.D.F., and P.L. A.M. contributed resources and software for image analysis. E.G.P., K.B.M., M.D.F., and P.L. led cohort development at UPenn, M.S. led cohort development at UH, and C.C., L.B., and A.P. contributed to cohort development at OSU. Resources, support, and technical oversight for digital slide set generation were contributed by A.P., M.D.F., K.B.M., M.S., and A.M. M.D.F., M.S., and C.C. contributed to re-grading efforts, as well as image quality and suitability for grading validation. Data analysis and interpretation were led by E.G.P. and S.A., with contributions from all other authors. All authors contributed to reviewing and revising the manuscript and approved the final version.

Ethics

The study was approved by the University of Pennsylvania institutional review board #7, protocol #830123, with waiver-of-consent authorized by 45 CFR 46.116(d) and 45 CFR 164.512(i), respectively (exemption 4). Given the largely de-identified and retrospective nature of this work, and the lack of present clinical implications from the potential diagnoses made on historical samples, it was neither reasonable nor appropriate to involve patients in the research or inform them of the results at its completion.

Conflict of interest: E.P. serves as a scientific consultant/advisory board member for CareDx and holds a patent related to the present work along with A.M., M.D.F., K.B.M., and A.J. M.D.F. is an equity holder and has technology licenced to both Elucid Bioimaging and Inspirata Inc., is a scientific advisory consultant for Inspirata Inc. and sits on its scientific advisory board, and is also a consultant for Phillips Healthcare, XFIN, and Virbio. K.B.M. serves as a scientific consultant/advisory board member for MyoKardia and Pfizer and he has had recent sponsored research agreements from Thoratec Corporation, Merck, Glaxo-Smith-Kline, and Sanofi-Aventis. A.M. is an equity holder in Elucid Bioimaging and in Inspirata Inc. In addition, he has served as a scientific advisory board member for Inspirata Inc, Astrazeneca, Bristol Meyers-Squibb, and Merck. Currently, he serves on the advisory board of Aiforia Inc. He also has sponsored research agreements with Philips and Bristol Meyers-Squibb. His technology has been licenced to Elucid Bioimaging. He is also involved in a NIH U24 grant with PathCore Inc, and three different R01 grants with Inspirata Inc. A.P. has nothing to disclose. C.C. has nothing to disclose. L.B. has nothing to disclose. P.L. has nothing to disclose. M.S. has nothing to disclose. S.A.-E. has nothing to disclose. S.A. has nothing to disclose.

Data availability

De-identified, digitized study histology slides can be made available in a limited capacity. Practical limitations related to the large file size of whole-slide digital images (often several gigabytes per slide) preclude complete cloud-based sharing of study image-sets, though representative images and tiles can be shared in a more complete fashion. Upon request, whole-image-sets could be shared with interested collaborators, though this will be made on a case-by-case basis due to the challenges and expense involved. Specific image analysis methods have already been thoroughly disclosed with appropriate citations and descriptions in the body of this manuscript, as transparency of methodology is a high priority. We have provided access to our source code at https://github.com/sarayar/CACHE-Grader.

References

1. Eisen HJ, Tuzcu EM, Dorent R, Kobashigawa J, Mancini D, Valantine-von Kaeppler HA, Starling RC, Sorensen K, Hummel M, Lind JM, Abeywickrama KH, Bernhardt P. Everolimus for the prevention of allograft rejection and vasculopathy in cardiac-transplant recipients. *N Engl J Med* 2003;**349**:847–858.

2. Kobashigawa JA, Miller LW, Russell SD, Ewald GA, Zucker MJ, Goldberg LR, Eisen HJ, Salm K, Tolzman D, Gao J, Fitzsimmons W, First R. Tacrolimus with mycophenolate mofetil (MMF) or sirolimus vs. cyclosporine with MMF in cardiac transplant patients: 1-year report. *Am J Transplant* 2006;**6**:1377–1386.
3. Patel JK, Kobashigawa JA. Should we be doing routine biopsy after heart transplantation in a new era of anti-rejection? *Curr Opin Cardiol* 2006;**21**:127–131.
4. Costanzo MR, Dipchand A, Starling R, Anderson A, Chan M, Desai S, Fedson S, Fisher P, Gonzales-Stawinski G, Martinelli L, McGiffin D, Smith J, Taylor D, Meiser B, Webber S, Baran D, Carboni M, Dengler T, Feldman D, Frigerio M, Kfoury A, Kim D, Kobashigawa J, Shullo M, Stehlik J, Teuteberg J, Uber P, Zuckermann A, Hunt S, Burch M, Bhat G, Canter C, Chinnock R, Crespo-Leiro M, Delgado R, Dobbels F, Grady K, Kao W, Lamour J, Parry G, Patel J, Pini D, Towbin J, Wolfel G, Delgado D, Eisen H, Goldberg L, Hosenpud J, Johnson M, Keogh A, Lewis C, O'Connell J, Rogers J, Ross H, Russell S, Vanhaecke J; International Society of Heart and Lung Transplantation Guidelines. The International Society of Heart and Lung Transplantation Guidelines for the care of heart transplant recipients. *J Heart Lung Transplant* 2010;**29**:914–956.
5. Billingham ME, Cary NR, Hammond ME, Kemnitz J, Marboe C, McCallister HA, Snovar DC, Winters GL, Zerbe A. A working formulation for the standardization of nomenclature in the diagnosis of heart and lung rejection: heart Rejection Study Group. The International Society for Heart Transplantation. *J Heart Transplant* 1990;**9**:587–593.
6. Hamilton D, *A History of Organ Transplantation: Ancient Legends to Modern Practice*. Pittsburgh, PA, USA: University of Pittsburgh Press; 2012.
7. Stewart S, Winters GL, Fishbein MC, Tazelaar HD, Kobashigawa J, Abrams J, Andersen CB, Angelini A, Berry GJ, Burke MM, Demetris AJ, Hammond E, Itescu S, Marboe CC, McManus B, Reed EF, Reinsmoen NL, Rodriguez ER, Rose AG, Rose M, Suciu-Focia N, Billingham ZA. Revision of the 1990 working formulation for the standardization of nomenclature in the diagnosis of heart rejection. *J Heart Lung Transplant* 2005;**24**:1710–1720.
8. Angelini A, Andersen CB, Bartoloni G, Black F, Bishop P, Doran H, Fedrigo M, Fries JW, Goddard M, Goebel H, Neil D, Leone O, Marzullo A, Ortmann M, Paraf F, Rotman S, Turhan N, Bruneval P, Frigo AC, Grigoletto F, Gasparetto A, Mencarelli R, Thiene G, Burke M. A web-based pilot study of inter-pathologist reproducibility using the ISHLT 2004 working formulation for biopsy diagnosis of cardiac allograft rejection: the European experience. *J Heart Lung Transplant* 2011;**30**:1214–1220.
9. Crespo-Leiro MG, Zuckermann A, Bara C, Mohacsi P, Schulz U, Boyle A, Ross HJ, Parameshwar J, Zakliczynski M, Fiocchi R, Stypmann J, Hoefer D, Lehmkuhl H, Deng MC, Leprince P, Berry G, Marboe CC, Stewart S, Tazelaar HD, Baron HM, Coleman IC, and Vanhaecke J. Concordance among pathologists in the second Cardiac Allograft Rejection Gene Expression Observational Study (CARGO II). *Transplantation* 2012;**94**:1172–1177.
10. Cruz-Roa A, Gilmore H, Basavanhally A, Feldman M, Ganesan S, Shih NNC, Tomaszewski J, Gonzalez FA, Madabhushi A. Accurate and reproducible invasive breast cancer detection in whole-slide images: a deep learning approach for quantifying tumor extent. *Sci Rep* 2017;**7**:46450.
11. Arevalo J, Cruz-Roa A, Arias V, Romero E, Gonzalez FA. An unsupervised feature learning framework for basal cell carcinoma image analysis. *Artif Intell Med* 2015;**64**:131–145.
12. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams T, Cuadros J, Kim R, Raman R, Nelson PC, Mega JL, Webster DR. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016; **316**:2402–2410.
13. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;**542**:115–118.
14. Golugula A, Lee G, Master SR, Feldman MD, Tomaszewski JE, and, Madabhushi A. Supervised regularized canonical correlation analysis: integrating histologic and proteomic data for predicting biochemical failures. *Conf Proc IEEE Eng Med Biol Soc* 2011;**2011**:6434–6437.
15. Leo P, Elliott R, Shih NNC, Gupta S, Feldman M, and, Madabhushi A. Stable and discriminating features are predictive of cancer presence and Gleason grade in radical prostatectomy specimens: a multi-site study. *Sci Rep* 2018;**8**:14918.
16. Lu C, Romo-Bucheli D, Wang X, Janowczyk A, Ganesan S, Gilmore H, Rimm D, and, Madabhushi A. Nuclear shape and orientation features from H&E images predict survival in early-stage estrogen receptor-positive breast cancers. *Lab Invest* 2018;**98**:1438–1448.
17. Lu C, Wang X, Prasanna P, Corredor G, Sedor G, Bera K, Velcheti V, and, Madabhushi A. Feature Driven Local Cell Graph (FeDeG): predicting overall survival in early stage lung cancer. *Lect Notes Comput Sci* 2018;**11071**:407–416.
18. Patil PD, Bera K, Vaidya P, Prasanna P, Khunger M, Khunger A, Velcheti V, and, Madabhushi A. Correlation of radiomic features with PD-L1 expression in early stage non-small cell lung cancer (ES-NSCLC) to predict recurrence and overall survival (OS). *J Clin Oncol* 2018;**36**:e24247–e24247.
19. Wang X, Janowczyk A, Zhou Y, Thawani R, Fu P, Schalper K, Velcheti V, and Madabhushi A. Prediction of recurrence in early stage non-small cell lung cancer using computer extracted nuclear features from digital H&E images. *Sci Rep* 2017;**7**:13543.
20. Corredor G, Wang X, Zhou Y, Lu C, Fu P, Syrigos K, Rimm DL, Yang M, Romero E, Schalper KA, Velcheti V, and, Madabhushi A. Spatial architecture and arrangement of tumor-infiltrating lymphocytes for predicting likelihood of recurrence in early-stage non-small cell lung cancer. *Clin Cancer Res* 2019;**25**: 1526–1534.
21. Jeffrey J, Nirschl AJ, Eliot G, Peyster Renee F, Kenneth B, Margulies Michael D, Feldman Anant, M. Deep learning tissue segmentation in cardiac histopathology images. In: SK Zhou, H Greenspan, D Shen, eds. *Deep Learning for Medical Image Analysis*; Amsterdam, NL: Elsevier 2017 pp. 179–195.
22. Nirschl JJ, Janowczyk A, Peyster EG, Frank R, Margulies KB, Feldman MD, and Madabhushi AA. Deep-learning classifier identifies patients with clinical heart failure using whole-slide images of H&E tissue. *PLoS One* 2018;**13**: e0192726.
23. Dooley AE, Tong L, Deshpande SR, Wang MD. Prediction of heart transplant rejection using histopathological whole-slide imaging. *IEEE EMBS Int Conf Biomed Health Inform* 2018. doi:10.1109/bhi.2018.8333416.
24. Duong VHJ, Fedrigo M, Fishbein GA, Leone O, Neil D, Marboe C, Peyster E, von D. T J, Loupy A, Mengel M, Revelo MP, Adam B, Bruneval P, Angelini A, Miller DV, Berry GJ. The XVth Banff Conference on Allograft Pathology the Banff Workshop Heart Report: improving the diagnostic yield from endomyocardial biopsies and Quilty effect revisited. *Am J Transplant* 2020;**20**:3308–3318.
25. Janowczyk A, Madabhushi A. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *J Pathol Inform* 2016;**7**:29
26. Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access* 2018;**6**:52138–52160.
27. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, Ferrero E, Agapow P-M, Zietz M, Hoffman MM, Xie W, Rosen GL, Lengerich BJ, Israeli J, Lanchantin J, Woloszynek S, Carpenter AE, Shrikumar A, Xu J, Cofer EM, Lavender CA, Turaga SC, Alexandari AM, Lu Z, Harris DJ, DeCaprio D, Qi Y, Kundaje A, Peng Y, Wiley LK, Segler MHS, Boca SM, Swamidass SJ, Huang A, Gitter A, Greene CS. Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface* 2018;**15**:20170387.
28. Bera K, Schalper KA, Rimm DL, Velcheti V, and, Madabhushi A. Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. *Nat Rev Clin Oncol* 2019;**16**:703–715.
29. Jiang Y, Bosch N, Baker R, Paquette L, Ocumpaugh J, Andres JMAL, Moorè AL, Biswas G. Expert feature-engineering vs. deep neural networks: which is better for sensor-free affect detection? *Lect Notes Comput Sci* 2018;**10947**:198–211.
30. Janowczyk A, Zuo R, Gilmore H, Feldman M, and, Madabhushi A. HistoQC: an open-source quality control tool for digital pathology slides. *JCO Clin Cancer Inform* 2019;**3**:1–7.
31. Ruifrok A, and Johnston D, Ruifrok AC, Johnston DA. Quantification of histochemical staining by color deconvolution. *Anal Quant Cytol Histol* **23**:291–299.
32. MacQueen JB. Some methods for classification and analysis of multivariate observations. *Proc Fifth Berkeley Symp Math Stat Probab* 1967;**1**:281–297.
33. Cawley GC, Talbot NLC. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J Mach Learn Res* 2010;**11**:2079–2107.
34. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960; **20**:37–46.
35. Bartko JJ. The intraclass correlation coefficient as a measure of reliability. *Psychol Rep* 1966;**19**:3–11.
36. Yuan Y. Modelling the spatial heterogeneity and molecular correlates of lymphocytic infiltration in triple-negative breast cancer. *J R Soc Interface* 2015;**12**: 20141153.
37. Le H, Gupta R, Hou L, Abousamra S, Fassler D, Torre-Healy L, Moffitt RA, Kurc T, Samaras D, Batiste R, Zhao T, Rao A, Van Dyke AL, Sharma A, Bremer E, Almeida JS, Saltz J. Utilizing automated breast cancer detection to identify spatial distributions of tumor-infiltrating lymphocytes in invasive breast cancer. *Am J Pathol* 2020;**190**:1491–1504.
38. Saltz J, Gupta R, Hou L, Kurc T, Singh P, Nguyen V, Samaras D, Shroyer KR, Zhao T, Batiste R, Van Arnam J, Shmulevich I, Rao AUK, Lazar AJ, Sharma A, Thorsson V. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell Rep* 2018;**23**: 181–193.e7.
39. Ström P, Kartasalo K, Olsson H, Solorzano L, Delahunt B, Berney DM, Bostwick DG, Evans AJ, Grignon DJ, Humphrey PA, Iczkowski KA, Kench JG, Kristiansen G, van der Kwast TH, Leite KRM, McKenney JK, Oxley J, Pan CC, Samaratunga

H, Srigley JR, Takahashi H, Tsuzuki T, Varma M, Zhou M, Lindberg J, Lindskog C, Ruusuvuori P, Wählby C, Grönberg H, Rantalainen M, Egevad L, Eklund M. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. *Lancet Oncol* 2020;**21**:222–232.

40. Bulten W, Pinckaers H, van Boven H, Vink R, de Bel T, van Ginneken B, van der Laak J, Hulsbergen-van de Kaa C, Litjens G. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *Lancet Oncol* 2020;**21**:233–241.
41. Marboe CC, Billingham M, Eisen H, Deng MC, Baron H, Mehra M, Hunt S, Wohlgemuth J, Mahmood I, Prentice J, Berry G. Nodular endocardial infiltrates (Quilty lesions) cause significant variability in diagnosis of ISHLT Grade 2 and 3A rejection in cardiac allograft recipients. *J Heart Lung Transplant* 2005;**24**:S219–26.
42. Rodriguez ER. The pathology of heart transplant biopsy specimens: revisiting the 1990 ISHLT working formulation. *J Heart Lung Transplant* 2003;**22**:3–15.
43. Tan CD, Baldwin WM, 3rd, Rodriguez ER. Update on cardiac transplantation pathology. *Arch Pathol Lab Med* 2007;**131**:1169–1191.
44. Aeffner F, Wilson K, Martin NT, Black JC, Hendriks CLL, Bolon B, Rudmann DG, Gianani R, Koegler SR, Krueger J, Young GD. The gold standard paradox in digital image analysis: manual versus automated scoring as ground truth. *Arch Pathol Lab Med* 2017;**141**:1267–1275.
45. Laurinavicius A, Laurinaviciene A, Dasevicius D, Elie N, Plancoulaine B, Bor C, Herlin P. Digital image analysis in pathology: benefits and obligation. *Anal Cell Pathol (Amst)* 2012;**35**:75–78.
46. Colvin MM, Cook JL, Chang P, Francis G, Hsu DT, Kiernan MS, Kobashigawa JA, Lindenfeld J, Masri SC, Miller D, O'Connell J, Rodriguez ER, Rosengard B, Self S, White-Williams C, Zeevi A. Antibody-mediated rejection in cardiac transplantation: emerging knowledge in diagnosis and management. *Circulation* 2015;**131**:1608–1639.