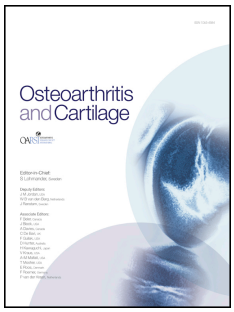
Journal Pre-proof



Deep learning enables the automation of grading histological tissue engineered cartilage images for quality control standardization

Laura Power, Lina Acevedo, Rikiya Yamashita, Daniel Rubin, Ivan Martin, Andrea Barbero

PII: S1063-4584(21)00004-2

DOI: https://doi.org/10.1016/j.joca.2020.12.018

Reference: YJOCA 4768

To appear in: *Osteoarthritis and Cartilage*

Received Date: 26 May 2020

Revised Date: 22 December 2020

Accepted Date: 28 December 2020

Please cite this article as: Power L, Acevedo L, Yamashita R, Rubin D, Martin I, Barbero A, Deep learning enables the automation of grading histological tissue engineered cartilage images for quality control standardization, *Osteoarthritis and Cartilage*, https://doi.org/10.1016/j.joca.2020.12.018.

Deep learning enables the automation of grading histological tissue engineered cartilage images for quality control standardization

Laura Power[1,2], Lina Acevedo[2], Rikiya Yamashita[3], Daniel Rubin[3], Ivan Martin[1,2*], Andrea Barbero[2]

[1] Department of Biomedical Engineering, University of Basel

[2] Department of Biomedicine, University Hospital Basel, University of Basel

[3] Department of Biomedical Data Science, Stanford University School of Medicine

* Corresponding author: ivan.martin@usb.ch, +41 61 265 23 84, University Hospital Basel, Hebelstrasse 20, CH-4031 Basel

laura.power@unibas.ch, linamarcelacevedo@gmail.com, rikiya@stanford.edu, rubin@stanford.edu, ivan.martin@usb.ch, andrea.barbero@usb.ch

Abstract

Objective

To automate the grading of histological images of engineered cartilage tissues using deep learning.

Methods

Cartilaginous tissues were engineered from various cell sources. Safranin O and fast green stained histological images of the tissues were graded for chondrogenic quality according to the Modified

Bern Score, which ranks images on a scale from zero to six according to the intensity of staining and cell morphology. The whole images were tiled, and the tiles were graded by two experts and grouped into four categories with the following grades: 0, 1-2, 3-4, and 5-6. Deep learning was used to train models to classify images into these histological score groups. Finally, the tile grades per donor were averaged. The root mean square errors (RMSEs) were calculated between each user and the model.

Results

Transfer learning using a pretrained DenseNet model was selected. The RMSEs of the model predictions and 95% confidence intervals were 0.49 (0.37, 0.61) and 0.78 (0.57, 0.99) for each user, which was in the same range as the inter-user RMSE of 0.71 (0.51, 0.93).

Conclusion

Using supervised deep learning, we could automate the scoring of histological images of engineered cartilage and achieve results with errors comparable to inter-user error. Thus, the model could enable the automation and standardization of assessments currently used for experimental studies as well as release criteria that ensure the quality of manufactured clinical grafts and compliance with regulatory requirements.

Keywords

Machine learning, transfer learning, convolutional neural networks, quality controls, regenerative medicine, histological score

Running title

DL for grading eng cartilage images

Introduction

Large cartilage defects do not have the capacity to regenerate in adults, and currently available treatments have not yet demonstrated predictable long-term efficacy[1]. Thus, new treatment options are required and being investigated[2–5]. One promising method is the implantation of autologous nasal chondrocyte-derived engineered tissue, which has been shown to be a safe and feasible method for treating knee-cartilage defects[6]. A phase II clinical trial is currently ongoing to test the efficacy of this treatment (BIO-CHIP: http://biochip-h2020.eu/). Briefly, nasal chondrocytes are isolated from the nasal septum, expanded, and seeded onto a collagen I/III scaffold. The resulting constructs are then cultured in chondrogenic condition, allowing the cells to produce their own cartilage matrix before implantation in the knee cartilage defect.

Release criteria must be developed for new therapies that assess a product to be of sufficient quality to perform its hypothesized mode of action[7]. For nasal chondrocyte-derived tissue engineered cartilage, the hypothesized mode of action is to fill defects with a mature cartilage-like matrix produced by the chondrocytes, thus restoring the functions of the knee and leading to increased mobility, decreased pain, and improved quality of life. The current release criteria for the BIO-CHIP study assesses the maturation of the engineered tissue by scoring histological images with a modified version of the Bern score, i.e., a well established method for grading the chondrogenicity of engineered cartilage[8].

Manual scoring systems have disadvantages that include low user agreement, subjectivity, potential for bias, and time consumption[9,10]. Yet, there is currently no automated method for grading the chondrogenicity of engineered tissues, only qualitative manual scoring systems[8,11]. Today, deep learning methods are increasingly being used for biomedical image analysis and are widely available to researchers. Deep learning has already been applied for tumor grading and classification[12,13] and grading osteoarthritis with micro-computed tomography[14].

Transfer learning is currently a very popular method being applied to new biomedical image analysis applications[15–18]. Transfer learning leverages deep learning models that have been pretrained with large datasets of a variety of images and can be fine-tuned with a smaller set of images. A pre-trained network can extract features from a small dataset that are then used to classify images[19], which is especially useful for developing deep learning models for medical applications, where datasets are relatively small.

In this study, with the final goal of automatically assessing the quality of cartilage grafts, we investigated whether deep learning can be used for grading histological images of tissue engineered cartilage.

Methods

Ethical approval

All human samples were collected with informed consent given by the involved individuals and in accordance with the cantonal ethical authority of Basel (Ethikkomission Nordwest- und Zentralschweiz; Ref.# 78/07) or the clinical trial (ClinicalTrials.gov, number NCT02673905).

Engineered cartilage

Chondrogenic micromasses and pellets

Nasal chondrocytes (NC), articular chondrocytes (AC), and mesenchymal stromal cells derived from bone marrow (BMSCs) and adipose tissue (ASCs) were expanded in monolayer and then cultured in micromass or pellet culture as previously described[20–23] using different culture media. More details on the chondrogenic culture protocols are provided in the supplemental materials.

Clinical engineered cartilage grafts

NCs were isolated from nasal septal cartilage biopsies, expanded, and cultured on collagen type I/III membranes (Chondro-Gide, Geistlich Pharma AG). Grafts for clinical use (here referred to as *clinical* grafts to distinguish them from experimental samples generated in the lab) were produced at the GMP facility at the University Hospital Basel according to standard operating procedures under a quality management system as previously described[6]. The clinical samples were collected from 3 female and 15 male patients with an average age of 37 (from 23 to 49). More details are provided in the supplemental materials.

Histological analysis

All samples were fixed in 4% formalin, embedded in paraffin, and sectioned to 5 µm thickness. Safranin O staining was performed with safranin O for glycosaminoglycans (GAG), fast green for collagen, and hematoxylin as a nuclear counterstaining as previously described[10].

Images were taken with the following microscopes: (1) Nikon upright Ni microscope with a Prior slide loader and a Nikon Ds-Fi3 camera and a CFI Plan Apo Lambda 20x objective (NA 0.75), (2) Nikon Ti2 microscope with a Nikon DS-Ri2 camera and a CFI Plan Apo Lambda 20x objective (NA 0.75), and (3) Olympus IX83 microscope with an Olympus DP80 camera and a LUCPlanFL N 20x objective (NA 0.45). The resolution of the images ranged from 0.24 to 0.5 µm/px. One or more images were taken of cartilage engineered from each donor. Evenly spread, non-overlapping areas of at least 300 x 300 px were extracted from each image, hereafter referred to as *tiles*, using a custom macro in Fiji/ImageJ (https://imagej.net/Fiji). Tiles were manually excluded from the datasets if they contained parts of the slide background or histological artifacts such as folded or torn tissues.

Modified Bern Score

Histological scoring via the Modified Bern Score (MBS) was performed on safranin O-stained histological images as previously described[21,23], adapted from Grogan et al.[10]. The MBS has two rating parameters, safranin O staining intensity and cell morphology that each receive a score between 0 and 3 (Table 1). The two values were summed together resulting in a maximum possible MBS of 6.

[Place Table 1 here]

In this manuscript, histological images were graded and grouped into four categories: *MBS 0*, *MBS 1-2*, *MBS 3-4*, and *MBS 5-6,* in order to classify images into groups, rather than handling scores on a continuous scale from zero to six. The general description of the quality of the engineered tissues in these four groups are as follows. MBS 0: no chondrogenesis, MBS 1-2: some cartilage attributes, MBS 3-4: moderate chondrogenesis, and MBS 5-6: Good to excellent chondrogenesis. Representative images of engineered cartilage for each category are displayed in Fig. 1.

153

154 [Place Fig. 1 here]

155

156 Clinical grafts were produced using Chondro-Gide, which is a bilayer graft. Cells were seeded on the
157 top permeable layer of the membrane where they can produce their own matrix during
158 chondrogenic culture. The bottom layer of the scaffold is impermeable to cells and provides
159 mechanical support to the construct after implantation. Only the top cell-laden layer was graded
160 when assessing the chondrogenicity of these grafts in the context of the clinical trial (Fig. S1). Tiles
161 containing the cell-free layer of the scaffold were manually discarded. The overall MBS per clinical
162 trial patient in this manuscript was calculated with Eqn. 1, where n is the number of tiles in each
163 class.

164

165 (Eqn. 1)

$$MBS_{patient} = \frac{(MBS\ 0)_n * 0 + (MBS\ 1\text{-}2)_n * 2 + (MBS\ 3\text{-}4)_n * 4 + (MBS\ 5\text{-}6)_n * 6}{(MBS\ 0)_n + (MBS\ 1\text{-}2)_n + (MBS\ 3\text{-}4)_n + (MBS\ 5\text{-}6)_n}$$

166

167 The effect of rounding the individual tile grades when dividing them into four classes was assessed
168 by comparing unrounded user grades and the result from Eqn. 1. A user graded the individual clinical
169 graft tiles on a scale from zero to six and the average MBS per patient was calculated, as is done in
170 the BIO-CHIP clinical trial. The tiles from the clinical grafts were then divided into the four classes
171 introduced in this manuscript using the common half round up method (e.g., 4.5 rounds to 5) and
172 then the average MBS per patient was calculated using Eqn. 1. An overall patient MBS grade ≥ 3 is
173 the threshold for a clinical graft to pass release criteria whereas a grade < 3 fails the release criteria
174 in the clinical trial, BIO-CHIP.

175

Dataset

The samples used for training and testing of the models are listed in Table 2. The training and validation images were graded by an expert user (user 1) and randomly split into about 80 and 20%, respectively, while ensuring an even distribution of images from each group, i.e., class, in the validation dataset.

The test data were obtained from nonoverlapping donors and experiments and included the clinical grafts produced for patients in the clinical trial (BIO-CHIP). The test samples were derived from 34 independent donors and individually scored by two experts (user 1 and 2).

[Place Table 2 here]

Model development

Python version 3.7.4 and the deep learning framework, PyTorch[24], were used to train and test models in this manuscript. Other Python libraries used were os, time, Matplotlib, the Python Imaging Library (PIL), and numpy. Calculations were performed at sciCORE (http://scicore.unibas.ch/) scientific computing center at University of Basel. We used the CentOS 7.5.1804 operating system, 64 GB RAM, and Intel Xeon CPU E5-2670 0 @ 2.60GHz. Training the neural network was performed using an Nvidia Titan X Pascal GPU and a CPU with 16 GB of RAM software allocation.

Workflow

The overall supervised learning workflow of classifying images into four quality groups and then taking the average score for all tiles derived from one donor is shown in Fig. 2.

[Place Fig. 2 here]

Data augmentation

The images in the training dataset were augmented using Python and PyTorch in order to increase the variability of the images presented to train the model. A series of data Image transforms from the Torchvision package[25] were applied to randomly resize (to ratios of 0.08 to 1.0 and random aspect ratios of 0.75 to 1.3) and crop the images to 224 x 224 pixels, slightly modify the colors with ColorJitter (brightness=0.1, contrast=0.1, hue=0.01), and randomly flip images, which were already inherently randomly orientated, horizontally (probability = 0.5) and vertically (probability = 0.5).

All the training, validation, and test images were normalized to the mean and standard deviation of the images in the training dataset, i.e., mean: 0.5894, 0.5352, 0.5669 and standard deviation: 0.0749, 0.0701, 0.0634.

The number of images per class in the training dataset were not evenly distributed (Table 2), therefore, a weighted random sampler was used during training to upsample images from the rarer classes. The per-class weights were the inverse of the number of images, i.e., 1/714, 1/549, 1/862, and 1/1205.

Model comparison

Table S1 lists all the deep learning models that were created and compared. A relatively small and simple convolutional neural network (CNN) was trained from scratch, similar to Bilaloglu et al.[26]. A summary of the model is included in the Supplementary Materials. Transfer learning was

implemented using MobileNet V2[27] and DenseNet161[28] architectures, which are pretrained on the ImageNet dataset (www.image-net.org) and available off-the-shelf for feature extraction and fine-tuning[19]. The fully-connected classification layers of each pretrained model were reshaped to predict the four classes in our dataset. During training, the trainable parameters in the original models were frozen, while only the parameters in the newly reshaped final layer were fine-tuned. Updating only the weights of the final layer allows for faster model training via transfer learning[18]. Details about the transfer learning models that were compared are provided in the supplementary materials.

Model training

Training was performed using the stochastic gradient descent algorithm by minimizing the cross entropy loss with a momentum factor of 0.9. A step-wise learning rate decay scheduler was used with a learning rate of 0.001, step size of 7, and gamma of 0.1 to improve the learning rate of the models[29]. The batch size was 32. The training was performed in 30 epochs, the trainable model parameters were saved from the epoch that achieved the highest validation accuracy. The model output was taken as the predicted class with the highest score.

Statistical analysis

The best performing model was determined using Cochran's Q test using the Mlxtend Python module[30] and the validation accuracy. Unless otherwise stated, all further statistical analyses were performed in R. The linear-weighted kappa statistic[31] was used to assess four-class classification, i.e., MBS 0, MBS 1-2, MBS 3-4, and MBS 5-6. Cohen's kappa was used to assess the pass/fail classification. Both kappa statistics were calculated with the psy package[32]. Confusion matrices were created using the Scikit-learn python module[33]. Receiver operating characteristic (ROC) and precision-recall curves, area under the curves (AUC) were calculated using the modEvA package[34]

and plotted using the multiROC package[35]. For the regression analysis of the final results, root mean square error (RMSE) was calculated with the ModelMetrics package[36]. Bootstrapping[37] was used to estimate the 95% confidence intervals of all the calculated statistics with the boot package[38] and 1000 resamples.

Model availability

The best model developed in this manuscript, which was evaluated with our test data, is available online at http://dx.doi.org/10.17632/wrdjkxhhs7.1

Visualize model decisions with Grad-CAM

To visually explain how the model predicted the label for each tile, gradient-weighted class activation mapping (Grad-CAM)[39] was performed by adapting the code from gradcam_plus_plus-pytorch[40]. The gradients flowing into the final convolutional layer for the predicted class were visualized to produce a coarse localization map to highlight the regions in the image that were important for selecting the final label or for each label.

Results

Calculate clinical graft scores

An example of the tiles scored for a clinical sample is displayed in Fig. 3. The overall patient MBS was calculated using Eqn. 1. The class distribution of tiles scored for each clinical patient graft is displayed in Fig. S2.

[Place Fig. 3 here]

Inter-user reliability

The linear-weighted kappa statistic between the two users was 0.46 (Fig. S3A). The normalized confusion matrix[36] shows how images in each group were classified differently by each user (Fig. S3B). The overall MBS per patient was calculated, and each donor passed the clinical release criteria with a grade ≥ 3 or failed with a grade < 3; the inter-user Cohen's kappa statistic for labeling a tile as pass or fail was 0.94 (Fig. 4).

Rounding error

Traditionally, the MBS allows flexibility for the user to give any score on a continuous scale of numbers. In this manuscript the scores are rounded so that they fit into four levels, according to Eqn. 1. The effect of this operation results in a slight loss of information (Fig. S4). Rounding increased the patient MBS by an average of 0.30 (standard deviation of 0.14).

Deep learning model comparison

CNNs with various architectures were trained and assessed with the validation data. The training loss and the validation accuracy of each model at each epoch is plotted in Fig. S5. All the models achieved their highest validation accuracies before 23 epochs, showing that 30 epochs were enough to train them to a comparable level and simultaneously reducing overfitting[41]. The best validation accuracies are displayed in Table S1. In the validation dataset, the number of images in each class were evenly distributed. Cochran's Q test was used to compare the performance of all seven models and did not reveal significant differences between them, with $p = 0.96$. Thus, the validation accuracy

of each model was used to compare them. The model with the best accuracy on the validation dataset was the transfer learning model using DenseNet and the newly trained, fully-connected linear classifier that mapped the output features from the pretrained model to the four classes in our dataset with a validation accuracy was 92.8%, thus this is the model we chose for analyzing the test data.

Deep learning predictions

Classification results

The test data were analyzed with the transfer learning model using DenseNet and a single fully-connected layer that was trained to map the features extracted by the pre-trained DenseNet to the four classes in our dataset. The model predictions were compared to the labels provided by each user for the classification part of the workflow (Fig. 2). The linear-weighted kappa statistic between each user and the model prediction for four classes was 0.64 and 0.47 for user 1 and 2, respectively (Fig. S3A). Normalized confusion matrices[42] show how the model predictions for each of the four classes compared with the labels provided by each user in Fig. S3C-D. Moreover, the model's ability to predict the user-provided label for images from each class are seen in ROC and precision-recall curves (Fig. S6A-B). The AUC for the micro-averaged ROC was 0.81 and 0.76, and the micro-averaged precision-recall AUC was 0.60 and 0.48 for user 1 and 2, respectively (Fig. S6C).

[Place Fig. 4 here]

Grad-CAM visualization

To visualize how the model predicted each label for the images, Grad-CAM was used to show what parts of each image were important for the network's decision (Fig. S7A). Incorrect model predictions could be visualized, and many images incorrectly labeled by the model showed the presence of tissues with varying quality within the same image (Fig. S7B-D), which also lead to user disagreement (Fig. S7E-F). When visualizing which parts of an image were activated for each label, the users could generally agree with the model on region-specific labels, based on the Grad-CAM visualization displayed in Fig. 5.

[Place Fig. 5 here]

Pass or fail prediction

The average MBS was calculated for each donor in the test dataset using Eqn. 1. The linear-weighted kappa statistic was 0.75 and 0.70 between the model and the labels provided by user 1 and 2, respectively (Fig. 4).

Overall MBS per donor

The final model evaluation step outlined in the workflow in Fig. 2 is to evaluate model's prediction of the overall MBS per individual patient or sample. The overall MBS may depend slightly on the number of tiles scored per patient or sample (Fig. S8). The average grade per donor based on the user labels were plotted against and the average grade predicted by the model (Fig. 6A). The model prediction RMSEs were 0.49 and 0.78 for user 1 and 2, respectively, which was in the same range as the inter-user error of 0.71 (Fig. 6B).

[Place Fig. 6 here]

Discussion

We showed for the first time that deep learning can be used to automatically grade images of tissue engineered cartilage according to a histological scoring system that is currently used to release grafts in a clinical setting. Transfer learning using a pretrained DenseNet model for feature extraction with a new fully-connected linear classification layer was trained to automatically grade histological images of engineered tissues that had RMSE in the range of the inter-user error.

The grading of histological images of nasal chondrocyte-derived tissue engineered cartilage products in an ongoing clinical trial (BIO-CHIP) is an important quality control method for characterization and standardization; therefore, clinical trial images were included in the test dataset in this manuscript. The model must be able to accurately predict the pass or fail threshold for clinical grafts, because it determines whether the graft can be released for implantation in the patient.

The automation of a comprehensive engineered cartilage scoring systems has not previously been reported. One component of a comprehensive scoring system, the staining intensity, can be calculated automatically without deep learning[9]. Color deconvolution could be used to split the colors based on images of tissue sections stained with only one color (i.e., only safranin O, only fast green, or only hematoxylin)[43], however, this method includes caveats related to background subtraction and staining variability. Attempts to automate the grading of the cell morphology category, however, have not provided promising results until now. With the recent availability of open source deep learning frameworks, it was natural to explore the use of this method to solve the problem of automating the grading of engineered cartilage.

382 The dataset used to train the deep learning model in this manuscript contains a good amount of
383 heterogeneity, with tissues engineered from three different cell types, i.e., NCs, ACs, and BMSCs.
384 The images of the histological tissue sections were taken with three different microscopes, resulting
385 in additional heterogeneity. In the future, the models could be retrained with images of engineered
386 cartilage with other microscope settings to further improve its generalizability. Nonetheless, thanks
387 to the variety of images used to train the model, it is already set up to be able to grade engineered
388 tissues generated under various experimental conditions. This generalizability was demonstrated by
389 the ability of the model to grade the quality of cartilage engineered from a fourth, unseen cell
390 source, i.e., ASCs. The automation of the Modified Bern Score with this model supports the
391 standardization of results across experimental conditions and tissue engineering laboratories or
392 manufacturing centers around the world.
393
394 To further improve this histological grading model, more fine-grained classes could be defined. In
395 this manuscript we binned the scores into four classes and envision that the number of groups could
396 be increased. This binning resulted in some rounding that needed to be performed in order to
397 calculate the per-donor scores; this rounding effect could be minimized in the future by adding more
398 bins. Although the effect of rounding was minimal here, it may be possible that this small rounding
399 effect would classify a sample as passing the set release criteria (score ≥ 3) when it would otherwise
400 be scored as failing (score < 3). An additional improvement during training of the model could be the
401 treatment of classes as ordinal[44].
402
403 The model had more agreement with the labels provided by user 1 than user 2 in the test dataset;
404 this is due to the fact that only user 1 provided ground truth labels for the training images. Providing
405 ground truth labels is time consuming, so the focus was placed on creating a well labeled test
406 dataset. Many incorrect model predictions could be explained by the presence of tissues with
407 varying quality within the same image, which in fact also lead to some user disagreement. This

highlights the need for an automated scoring system to increase standardization, and shows the dependence that deep learning has in this context on the accuracy of the user provided labels. In the future, a set of images should be thoroughly graded by multiple expert users, which will necessitate the discussion of more specific criteria for each class. This effort should focus on highly confident image labels rather than quantity of images, since we see that thanks to transfer learning, convolutional neural networks can be successfully trained for this purpose with just a few hundred images.

The training dataset in this manuscript consisted solely of engineered micromass pellets, so every tissue-containing image tile could be included. The clinical testing dataset, however, had only the parts of the images that contained the upper permeable layer of the scaffold. We had to decide on a cut-off boundary between the permeable and impermeable parts of each engineered tissue. This was complicated when some remaining collagen fibers from the scaffold extended into the upper cartilaginous portion of the mature graft and caused the model to misclassify some images. As with the current manual Modified Bern Scoring system, subjectivity is introduced to the scoring process when a user must decide which regions of the engineered grafts to score. Moreover, although the final grade per patient was estimated based on 34 independent samples, when calculating the linear-weighted kappa scores for the tiles, the assumption of independent donors was violated. In the future, the tiling part of the scoring process could be further automated, possibly with an attention-based model that more highly weights the most significant aspects of a whole slide image[45].

Clinical outcome data after two and five years will be collected from the patients in the ongoing clinical trial (BIO-CHIP) using the KOOS scoring system, where patients report scores on mobility, pain, ability to do sport, overall quality of life, etc.[46]. Once this clinical outcome data is available, the

histological images of the engineered grafts can again be reviewed, and correlations with the histological score and other features of the graft[21,47] can be investigated.

In conclusion, thanks to the recent advances in deep learning, it is now possible to automate the grading of histological images of engineered cartilage, resulting in faster readouts, reducing the need for pathologists with years of experience, and reducing scoring bias. An automated method to score images of engineered cartilage will certainly be of great interest to all researchers who investigate chondrogenesis. Moreover, the standardization and increased quantitation of quality controls in tissue engineering will allow us to more objectively assess grafts, providing us with richer information about the advanced therapy medicinal products (ATMPs) that are implanted in patients. The more accurate the characterization data of regenerative medicines is, the more knowledge we will have when analyzing the clinical outcome, which will allow for the improved treatment of patients.

Acknowledgements

We would like to thank Sandra Feliciano, Francine Wolf, Dr. Paola Occhetta, Mansoor Chaaban, and Laura Dönges for engineering the tissues; Dr. M. Adelaide Asnaghi, Majoska Berkelaar, and Evan Kotler for supporting image scoring; and Dr. Loïc Sauteur for the histomorphometry work. Calculations were performed at the sciCORE (http://scicore.unibas.ch/) scientific computing center at the University of Basel.

Authors contributions

The article was conceived and designed by LP and LA. LP drafted the article, assembled the data, and performed the statistical analyses. LP and LA prepared the dataset labels. Analysis and interpretation of the data was performed by LP, RY, and DR. All the authors critically revised the article for important intellectual content and provided final approval. LP takes full responsibility for the integrity of the work as a whole, from inception to finished article.

Role of funding source

This project has received funding from the European Union's Horizon 2020 research and innovation program under Grant Agreement No. 681103 (BIO-CHIP). This project was also supported by the Freiwillige Akademische Gesellschaft, Basel.

Conflicts of interest

The authors have no conflict of interests to disclose.

References

1. Brix MO, Stelzeneder D, Chiari C, Koller U, Nehrer S, Dorotka R, et al. Treatment of full-thickness chondral defects with hyalograft C in the Knee: Long-term results. Am J Sports Med. 2014;42(6):1426–32.
2. Gatenholm B, Lindahl C, Brittberg M, Simonsson S. Collagen 2A Type B Induction after 3D Bioprinting Chondrocytes In Situ into Osteoarthritic Chondral Tibial Lesion. Cartilage. 2020;
3. Brusalis CM, Greditzer HG, Fabricant PD, Stannard JP, Cook JL. BioCartilage augmentation of marrow stimulation procedures for cartilage defects of the knee: Two-year clinical outcomes.

Knee. 2020;27(5):1418–25.

4. Vahedi P, Hosainzadegan H, Brazvan B, Roshangar L, Shafaei H, Salimnejad R. Treatment of cartilage defects by Low-intensity pulsed ultrasound in a sheep model. Cell Tissue Bank. 2020;

5. Shah SS, Mithoefer K. Scientific Developments and Clinical Applications Utilizing Chondrons and Chondrocytes with Matrix for Cartilage Repair. Cartilage. 2020;

6. Mumme M, Barbero A, Miot S, Wixmerten A, Feliciano S, Wolf F, et al. Nasal chondrocyte-based engineered autologous cartilage tissue for repair of articular cartilage defects: an observational first-in-human trial. Lancet. 2016;388(10055):1985–94.

7. Bravery C, Carmen J, Fong T, Oprea W, Hoogendoorn K, Woda J, et al. Potency assay development for cellular therapy products: An ISCT* review of the requirements and experiences in the industry. Cytotherapy. 2013;15:9–19.

8. Rutgers M, van Pelt MJP, Dhert WJA, Creemers LB, Saris DBF. Evaluation of histological scoring systems for tissue-engineered, repaired and osteoarthritic cartilage. Osteoarthr Cartil. 2010;18(1):12–23.

9. O'Driscoll SW, Marx RG, Beaton DE, Miura Y, Gallay SH, Fitzsimmons JS. Validation of a simple histological-histochemical cartilage scoring system. Tissue Eng. 2001;7(3):313–20.

10. Grogan SP, Barbero A, Winkelmann V, Rieser F, Fitzsimmons JS, O'Driscoll S, et al. Visual Histological Grading System for the Evaluation of in Vitro-Generated Neocartilage. Tissue Eng. 2006;12(8):2141–9.

11. Maglio M, Brogini S, Pagani S, Giavaresi G, Tschon M. Current Trends in the Evaluation of Osteochondral Lesion Treatments: Histology, Histomorphometry, and Biomechanics in Preclinical Models. Biomed Res Int. 2019;4040236.

12. Ryu HS, Jin M-S, Park JH, Lee S, Cho J, Oh S, et al. Automated Gleason Scoring and Tumor Quantification in Prostate Core Needle Biopsy Images Using Deep Neural Networks and Its Comparison with Pathologist-Based Assessment. Cancers (Basel). 2019;11(12):1860.

13. Wang Y, Guan Q, Lao I, Wang L, Wu Y, Li D, et al. Using deep convolutional neural networks

for multi-classification of thyroid tumor by histopathology: a large-scale pilot study. Ann Transl Med. 2019;7(18):468–468.

14. Rytky SJO, Tiulpin A, Frondelius T, Finnilä MAJ, Karhula SS, Leino J, et al. Automating three-dimensional osteoarthritis histopathological grading of human osteochondral tissue using machine learning on contrast-enhanced micro-computed tomography. Osteoarthr Cartil. 2020;28(8):1133–44.

15. Saikia AR, Bora K, Mahanta LB, Das AK. Comparative assessment of CNN architectures for classification of breast FNAC images. Tissue Cell. 2019;57:8–14.

16. Gessert N, Bengs M, Wittig L, Drömann D, Keck T, Schlaefer A, et al. Deep transfer learning methods for colon cancer classification in confocal laser microscopy images. Int J Comput Assist Radiol Surg. 2019;14(11):1837–45.

17. Mazo C, Bernal J, Trujillo M, Alegre E. Transfer learning for classification of cardiovascular tissues in histological images. Comput Methods Programs Biomed. 2018;165:69–76.

18. Rivenson Y, Wang H, Wei Z, de Haan K, Zhang Y, Wu Y, et al. Virtual histological staining of unlabelled tissue-autofluorescence images via deep learning. Nat Biomed Eng. 2019;3(6):466–77.

19. Mormont R, Geurts P, Maree R. Comparison of deep transfer learning strategies for digital pathology. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. IEEE Computer Society; 2018. p. 2343–52.

20. Ghosh S, Spagnoli GC, Martin I, Ploegert S, Demougin P, Heberer M, et al. Three-dimensional culture of melanoma cells profoundly affects gene expression profile: A high density oligonucleotide array study. J Cell Physiol. 2005;204(2):522–31.

21. Asnaghi MA, Power LJ, Barbero A, Haug M, Köppl R, Wendt D, et al. Biomarker signatures of quality for engineering nasal chondrocyte-derived cartilage. Front Bioeng Biotechnol. 2020;8:283.

22. Osinga R, Di Maggio N, Todorov A, Allafi N, Barbero A, Laurent F, et al. Generation of a Bone

534 Organ by Human Adipose-Derived Stromal Cells Through Endochondral Ossification. Stem
535 Cells Transl Med. 2016;5(8):1090–7.
536 23. Lehoczky G, Wolf F, Mumme M, Gehmert S, Miot S, Haug M, et al. Intra-individual
537 comparison of human nasal chondrocytes and debrided knee chondrocytes: Relevance for
538 engineering autologous cartilage grafts. Clin Hemorheol Microcirc. 2019;74(1):67–78.
539 24. Paszke A, Gross S, Chintala S, Chanan G, Yang E, Facebook ZD, et al. Automatic differentiation
540 in PyTorch. In: 31st Conference on Neural Information Processing Systems (NIPS 2017). Long
541 Beach, CA; 2017.
542 25. Marcel S, Rodriguez Y. Torchvision the machine-vision package of torch. In: MM'10 -
543 Proceedings of the ACM Multimedia 2010 International Conference. New York, New York,
544 USA: ACM Press; 2010. p. 1485–8.
545 26. Bilaloglu S, Wu J, Fierro E, Sanchez RD, Ocampo PS, Razavian N, et al. Efficient pan-cancer
546 whole-slide image classification and outlier detection using convolutional neural networks.
547 bioRxiv. 2019;doi.org/10.1101/633123.
548 27. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C. MobileNetV2: Inverted Residuals and
549 Linear Bottlenecks. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2018;4510–20.
550 28. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely Connected Convolutional
551 Networks. Proc - 30th IEEE Conf Comput Vis Pattern Recognition, CVPR 2017. 2017;4700–8.
552 29. Subramanian V. Deep learning with PyTorch: a practical approach to building neural network
553 models using PyTorch. Birmingham UK: Packt Publishing; 2018.
554 30. Raschka S. MLxtend: Providing machine learning and data science utilities and extensions to
555 Python's scientific computing stack. J Open Source Softw. 2018;3(24):638.
556 31. Cohen J. Weighted kappa: Nominal scale agreement provision for scaled disagreement or
557 partial credit. Psychol Bull. 1968;70(4):213–20.
558 32. Falissard B. psy: Various procedures used in psychometry. R package version 1.1. [Internet].
559 2012. Available from: https://cran.r-project.org/web/packages/psy/index.html

33. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. 2012;

34. Barbosa AM, Real R, Muñoz A-R, Brown JA. New measures for assessing model equilibrium and prediction mismatch in species distribution models. Robertson M, editor. Divers Distrib. 2013;19(10):1333–8.

35. Wei R, Wang J, Jia W. multiROC: Calculating and Visualizing ROC and PR Curves Across Multi-Class Classifications. R package version 1.1.1 [Internet]. 2018. Available from: https://cran.r-project.org/web/packages/multiROC/index.html

36. Hunt T. ModelMetrics: Rapid Calculation of Model Metrics [Internet]. Comprehensive R Archive Network (CRAN); 2020. Available from: https://cran.r-project.org/package=ModelMetrics

37. Davison AC, Hinkley D V. Bootstrap Methods and Their Application. Cambridge: Cambridge University Press; 1997.

38. Canty A, Ripley B. boot: Bootstrap R (S-Plus) Functions. R package version 1.3-24 [Internet]. 2019. Available from: https://cran.r-project.org/web/packages/boot/

39. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. Int J Comput Vis. 2016;128(2):336–59.

40. gradcam_plus_plus-pytorch: A Simple pytorch implementation of GradCAM and GradCAM++ [Internet]. Available from: https://github.com/vickyliin/gradcam_plus_plus-pytorch

41. Prechelt L. Early Stopping — But When? In: Montavon G, Orr GB, Müller K, editors. Berlin, Heidelberg: Springer; 2012. p. 53–67.

42. Simske S. Meta-analytic design patterns. In: Meta-Analytics. Elsevier; 2019. p. 147–85.

43. Ruifrok AC, Johnston DA. Quantification of histochemical staining by color deconvolution. Anal Quant Cytol Histol. 2001;23(4):291–9.

44. Cheng J, Wang Z, Pollastri G. A neural network approach to ordinal regression. In:

Proceedings of the International Joint Conference on Neural Networks. 2008. p. 1279–84.

45. Momeni A, Thibault M, Gevaert O. Deep Recurrent Attention Models for Histopathological Image Analysis. bioRxiv. 2018;doi.org/10.1101/438341.

46. Roos EM, Roos HP, Lohmander LS, Ekdahl C, Beynnon BD. Knee Injury and Osteoarthritis Outcome Score (KOOS) - Development of a self-administered outcome measure. J Orthop Sports Phys Ther. 1998;28(2):88–96.

47. Power LJ, Wixmerten A, Wendt D, Barbero A, Martin I. Raman spectroscopy quality controls for GMP compliant manufacturing of tissue engineered cartilage. In: Proceedings Volume 10881, Imaging, Manipulation, and Analysis of Biomolecules, Cells, and Tissues XVII. San Francisco, CA: SPIE-Intl Soc Optical Eng; 2019. p. 108810F.

Figure legends

Fig. 1

Representative images of histological scores. Histological images were given a score (0-3) for the safranin O staining intensity and a score (0-3) for the cell morphology. The sum of these two categories is the Modified Bern Score (MBS), which grades the chondrogenicity of in vitro engineered cartilage on a scale of zero to six. Here, images were grouped into four categories: MBS 0, MBS 1-2, MBS 3-4, and MBS 5-6.

Fig. 2

Workflow. The workflow of supervised model training and testing is depicted. Whole images of engineered tissues from individual donors were tiled, labeled with one of four quality categories, and used to train various convolutional neural network architectures. The best model was selected and tested with an independent test dataset. For each patient or donor in the test dataset, the average histological score was calculated based on the user labels and labels predicted by the model and assessed with the root mean square error (RMSE).

Fig. 3

An engineered cartilage graft from a clinical patient. The tiles used for scoring were taken from the top chondrogenic layer of the graft, some of which are illustrated here with red boxes. The grades given by user 1 are displayed. The overall patient MBS was calculated based on the number of tiles in each class and using Eqn. 1. Here the overall patient MBS is 3.3 and the tiles are 122 x 122 µm.

622

623 Fig. 4

624

625 **Pass or fail results on the test dataset.** For each donor in the test dataset, pass or fail model
626 predictions vs. user labels and inter-user reliability assessed with the Cohen's kappa statistic and
627 95% confidence interval.

628

629 Fig. 5

630

631 **Grad-CAM visualizations.** One image that contains several grades of tissue engineered cartilage
632 quality was given regional labels by the users. The users agreed on the overall label MBS 3-4 while
633 the model predicted the label MBS 1-2. Grad-CAM visualized the regions of the image that were
634 highlighted for all four grading categories.

635

636 Fig. 6

637

638 **Model predictions of the overall patient MBS.** The overall histological Modified Bern Score for each
639 donor in the test dataset (n = 34) was calculated using Eqn. 1 with the labels predicted by the model
640 and the labels provided by each user. (A) The predicted patient grades plotted against the labels
641 provided by each user for the clinical grafts (patients) and experimental engineered tissues. (B) The
642 root mean square (RMSE) for each comparison was plotted along with the 95% confidence intervals.

643

Tables

Table 1

The Modified Bern Score assesses histological images of engineered cartilage based on the following two categories, each of which receives a score between zero and three that are then added together.

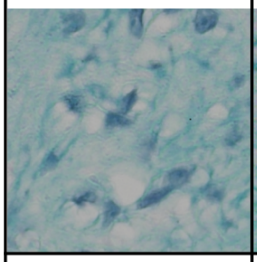
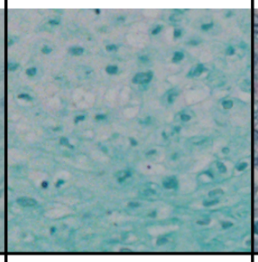
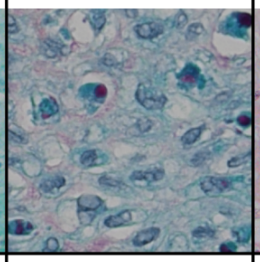
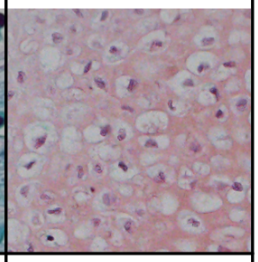
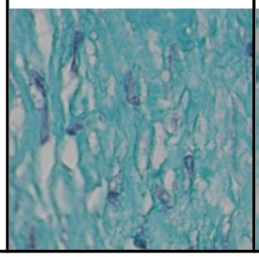
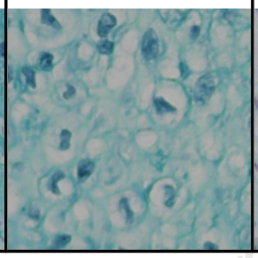
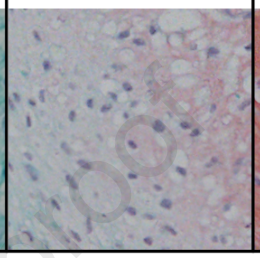
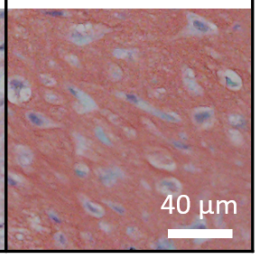
Scoring category	Score	Definition
Intensity of safranin O staining	0	No stain
	1	Weak staining
	2	Moderately even staining
	3	Even dark stain
Cell morphology	0	Condensed/necrotic/pycnotic bodies
	1	Spindle/fibrous
	2	Mixed spindle/fibrous with rounded chondrogenic morphology
	3	Majority rounded/chondrogenic

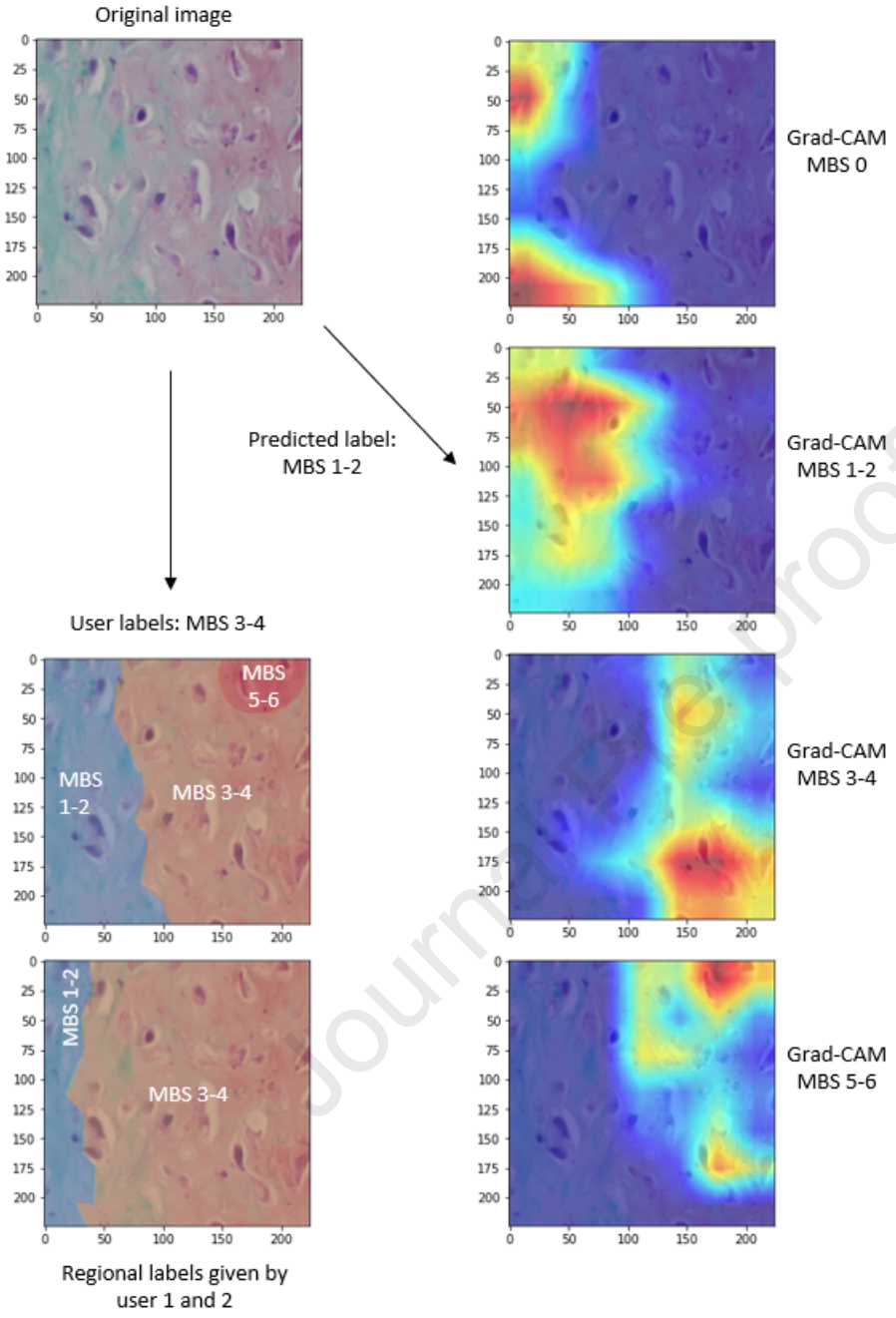
651 Table 2

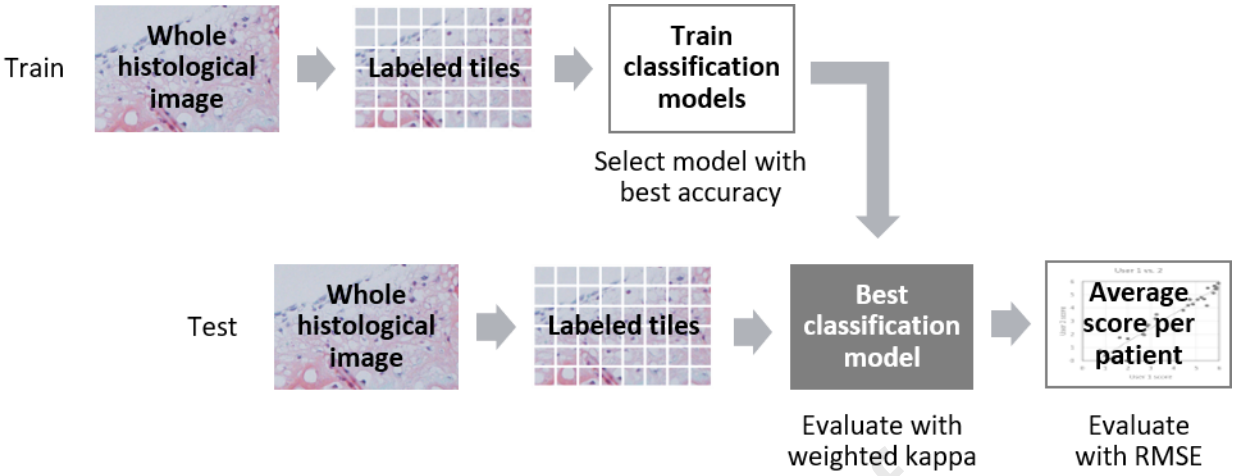
652 Samples used for training and testing the models. BMSC = bone-marrow derived mesenchymal
653 stromal cells, ASC = adipose tissue-derived stromal cells, NC = nasal chondrocytes, and AC = articular
654 chondrocytes. The labels according to user 1 are displayed.

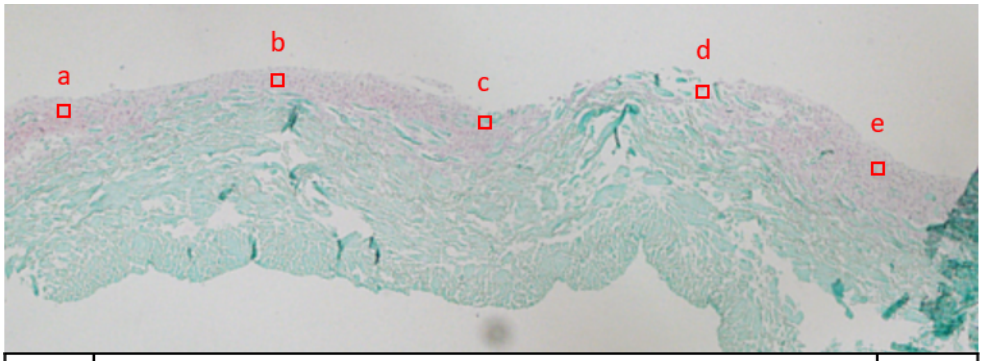
Dataset	Experimental or clinical samples	Cell type and number of donors	Microscope used	Graded by	Number of tiles	Number of tiles per category			
						MBS 0	MBS 1-2	MBS 3-4	MBS 5-6
Training	Experimental	7 BMSC, 13 NC, and 3 AC	Nikon Ni and Olympus IX83	User 1	3330	714	549	862	1205
Validation					600	150	150	150	150
Test	Experimental	5 AC, 4 NC, 2 BMSC, and 5 ASC	Nikon Ti2	User 1 and 2	383	124	160	65	34
	Clinical	18 NC	Olympus IX83		679	35	14	88	542
	Total test tiles (n = 34)				1062	159	174	153	576

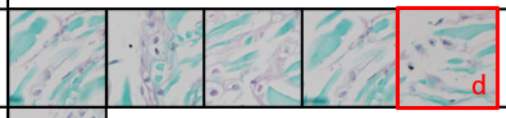
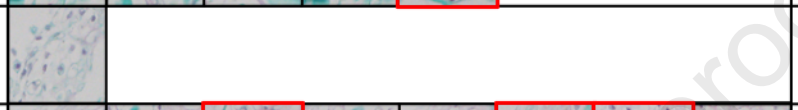
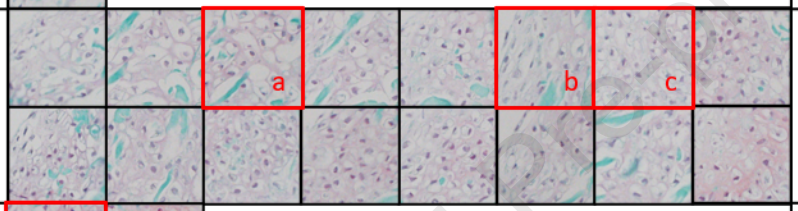
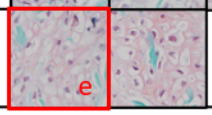
655

Modified Bern Score group	MBS 0	MBS 1-2	MBS 3-4	MBS 5-6
Safranin O staining intensity (0 to 3) + cell morphology (0 to 3) = Modified Bern Score	0 + 0 = 0	0 + 1 = 1	1 + 2 = 3	2 + 3 = 5
				
	0 + 0 = 0	0 + 2 = 2	2 + 2 = 4	3 + 3 = 6
				







MBS group	Tiles	Num. images
0		5
1-2		1
3-4		16
5-6		2