EARLY ONLINE RELEASE

Note: This article was posted on the Archives Web site as an Early Online Release. Early Online Release articles have been peer reviewed, copyedited, and reviewed by the authors. Additional changes or corrections may appear in these articles when they appear in a future print issue of the Archives. Early Online Release articles are citable by using the Digital Object Identifier (DOI), a unique number given to every article. The DOI will typically appear at the end of the abstract.

The DOI for this manuscript is doi: 10.5858/arpa.2020-0510-OA

The final published version of this manuscript will replace the Early Online Release version at the above DOI once it is available.
Histopathologic and Machine Deep Learning Criteria to Predict Lymphoma Transformation in Bone Marrow Biopsies

Lina Irshaid, MD; Jonathan Bleiberg, BS; Ethan Weinberger, BS; James Garritano, BS; Rory M. Shallis, MD; Jonathan Patsenker, BS; Ofir Lindenbaum, PhD; Yuval Kluger, PhD; Samuel G. Katz, MD, PhD; Mina L. Xu, MD

Objectives—To study the accuracy and reproducibility of a trained convolutional neural network in identifying LCT, in light of promising machine learning tools that may introduce greater objectivity to morphologic analysis.

Design.—We retrospectively identified patients who had a diagnosis of FL or CLL who had undergone bone marrow biopsies for the clinical question of LCT. We scored morphologic criteria and correlated results with clinical disease progression. In addition, whole slide scans were annotated into patches to train convolutional neural networks to discriminate between small and large tumor cells and to predict the patient’s probability of transformation.

Results.—Using morphologic examination, the proportion of large lymphoma cells (≥10% in FL and ≥30% in CLL), chromatin pattern, distinct nucleoli, and proliferation index were significantly correlated with LCT in FL and CLL. Compared to pathologist-derived estimates, machine generated quantification demonstrated better reproducibility and stronger correlation with final outcome data.

Conclusions.—These histologic findings may serve as indications of LCT in bone marrow biopsies. The pathologist—augmented with machine system appeared to be the most predictive, arguing for greater efforts to validate and implement these tools to further enhance physician practice.

(Arch Pathol Lab Med. doi: 10.5858/arpa.2020-0510-OA)

Accepted for publication February 19, 2021.

From the Departments of Pathology (Irshaid, Garritano, Patsenker, Kluger, Katz, Xu) and Internal Medicine (Shallis), Yale New Haven Hospital, Yale School of Medicine, New Haven, Connecticut; and the Program of Applied Mathematics, Yale University, New Haven, Connecticut (Bleiberg, Weinberger, Lindenbaum, Kluger). Irshaid and Bleiberg contributed equally to this work.

The authors have no relevant financial interest in the products or companies described in this article.

Preliminary data from this research was presented at United States and Canadian Academy of Pathology Annual Meeting; March 18, 2019; Washington, DC.

Corresponding author: Mina L. Xu, MD, Yale Medicine, 310 Cedar Street, PO Box 208023, New Haven, CT 06510 (email: mina.xu@yale.edu).

Arch Pathol Lab Med

Predictions of Lymphoma Transformation—Irshaid et al
Patients undergoing Richter transformation may also experience increasing lymphadenopathy, decreasing complete blood counts, and worsening of systemic symptoms. Although the diagnosis is usually made on lymph node or soft tissue biopsies, only 1 prior study, upon our literature review, has shown that counting the percentage of large cells in the marrow can be used as a surrogate. Such findings can be very helpful given that lymph node biopsies are invasive and risk missing the area of transformation. However, implementation of these findings is challenged by the difficulty of reaching histologic consensus using thresholds such as 7% large cells, which had been a proposal from the prior study investigating large-cell percentage in marrows involved by CLL.

In this study, we analyze whole slide images (WSIs) to extract relevant histopathologic information that could be analyzed with a convolutional neural network (CNN)–based approach. We apply this approach to a set of cases with WSI data to predict the probability of LCT in FL and CLL. In addition, pathologists blinded to the clinical determinants and outcome scored discrete morphologic categories in order to demonstrate the predictability of each variable for LCT. We compare these 2 methods of morphologic assessment, traditional pathologic criteria, and our deep learning algorithm to ascertain the utility of either alone and in combination.

**MATERIALS AND METHODS**

After fulfilling Institutional Review board requirements, we searched the pathology files for patients with a previous diagnosis of FL or CLL who had undergone a bone marrow biopsy for the clinical question of LCT. All patients identified were included in the study. In patients for whom multiple bone marrow biopsies were obtained to assess for LCT, only the initial biopsy was included in the study. Patients with nondefinitive FL or CLL diagnoses, or those with other associated bone marrow pathologies (eg, concomitant myeloid/lymphoid disease, myelodysplastic syndrome, or metastatic carcinoma) were excluded. For the remaining patients, the pathology archives were also searched for concomitant or immediately subsequent (within 3 weeks) lymph node or soft tissue biopsies.

**Histologic Scoring**

Two pathologists blinded to diagnostic report and outcome independently scored all hematoxylin-eosin–stained bone marrow biopsies for the percentage of marrow cellularity involved by lymphoma, the percentage of large cells within the lymphoma, the presence of highly atypical cells (0, absent; 1, present), the presence of irregular lymphoid aggregates (0, absent; 1, present), the presence of distinct nucleoli among tumor cells (0, absent; 1, present), and the chromatin pattern among tumor cells (0, condensed; 1, open). Additionally, reticulin and Ki-67 stains were used to assess degree of fibrosis (0, absent; 1, mild; 2, moderate; 3, marked) and proliferation index (% of lymphoma), respectively. Irregular lymphoid aggregates were defined as aggregates that did not conform to typical paratrabecular, nodular, or small interstitial clusters, and included coalescing and serpiginous aggregates (Figure 1). Highly atypical cells were defined as cells with marked irregularities of the nucleus, usually presenting with striking cytologic features (Figure 2).

Results were correlated with “final clinical transformation outcomes.” Ultimate evidence of LCT was based on concomitant or immediately subsequent lymph node/soft tissue biopsy, when available, or clinical decision to initiate aggressive chemotherapy. The frequency with which marrow microscopy is predictive of clinical transformation as well as the discordance of marrow with lymph node biopsies was assessed.

**Figure 1.** Irregular lymphoid aggregates in marrow examined for large-cell transformation. Bone marrow core biopsies demonstrating irregular lymphoid aggregates, including large interstitial aggregates (A), aggregates transitioning into a diffuse infiltrate (B), paratrabecular aggregates with a prominent interstitial component (C), and coalescing and serpiginous aggregates (D) (hematoxylin-eosin, original magnification ×200).
Clinical Data

Overall clinical data obtained included patient race, sex, age at initial lymphoma diagnosis and at suspicion of LCT, Eastern Cooperative Oncology Group Performance Status, human immunodeficiency virus (HIV) status, lactate dehydrogenase (LDH) at initial lymphoma diagnosis and at suspicion of LCT, B symptoms (defined as fever, night sweats, or weight loss), IPI score at LCT, Ann Arbor stage at initial lymphoma diagnosis and at suspicion of LCT, type of therapy, fluorodeoxyglucose-positron emission tomography/computed tomography response, clinical response, hematopoietic stem cell transplantation (autologous or allogeneic), relapse-free survival, and overall survival. Data obtained specifically for FL included Follicular Lymphoma IPI score, grade at diagnosis, BCL2/BCL6 protein expression, and BCL2/BCL6/C-MYC rearrangement status. Data obtained for CLL included Rai stage, β2-microglobulin levels at diagnosis and at suspicion for transformation, immunoglobulin heavy chain (IGHV) mutation status, presence of complex karyotype (defined as >3 aberrations) or del(17p), and flow cytometric expression of CD38 (<30% versus >30%) or ZAP70 (<20% versus >20%).

Machine Learning: Data Preparation

All FL and CLL cases were scanned at ×40 magnification using a high-resolution Aperio scanner (Aperio ScanScope CS, Aperio Technologies, Vista, California) and annotated with the digital pathology analysis software QuPath to define areas of maturing trilineage hematopoiesis, small-cell lymphoma and large-cell lymphoma (Figure 3). The total number of resolved pixels in the image files varied for each patient, ranging from 33 120 × 30 270 to 128 160 × 76 800 for the FL patients and 34 560 × 17 280 to 183 320 × 97 920 for the CLL patients. QuPath annotated regions were first used to create machine-derived estimates of relative large-cell and small-cell lymphoma prevalence by surface area, which were set aside for later statistical analysis.
Image data were then extracted from the QuPath regions of interest to serve as an input to a neural network–based image analysis algorithm. In particular, the QuPath Groovy scripting interface was used to extract the polygon coordinates defining the annotated regions of the WSIs. The resulting regions were then split into a grid of 128 × 128 pixel image patches and extracted as JPG files with the Python bindings of the OpenSlide library. In order to make the most efficient use of the limited data available, a “stride augmentation” method was employed (Figure 4), in which WSIs were resampled using several partially overlapping, offset grids, effectively augmenting the data set with image patches only partially overlapped to those extracted from previous grids. The number of patches extracted per patient varied based on the size of the regions of interest annotated by the pathologist. For FL, a mean of 3507 patches were extracted for each patient (standard deviation, 5191 patches), whereas for CLL, a mean of 8675 patches were extracted per patient (standard deviation, 9540 patches). Because of the large variability in available patch numbers per patient, a carefully designed sampling procedure was used to ensure that all patients were represented in training and validation (see Model Training and Architecture Specification).

After extraction from QuPath, the resulting image patches were preprocessed by a Python script that used a traditional histogram thresholding–based algorithm to determine which patches represented relevant tissue rather than white space or fat/bone cell regions. Patches were first cast to grayscale and rescaled so that all pixel values were between 0 and 1. The patches were then convolved with a Gaussian blur filter of radius 7 px. The blur filter served to smooth over the white space area present between cells. The pixel values were then thresholded at 0.9, with any pixel value greater than the threshold declared to be white space. This created a Boolean white space mask, from which the patchwise percentage of white space area could easily be calculated (Figure 5). Empirical trials found the algorithm to work best when patches with greater than 90% white space area were removed from consideration. These preprocessing steps yielded a data set of 128 × 128 pixel image patches for each patient, which could be used as input to a CNN-based machine learning algorithm.

Deep Learning Algorithm Development

The resulting data set of image patches was input to 2 separate CNN-based algorithms, the results from which were compared to determine the best method for determining the probability that a patient exhibits LCT. Cases without complete clinical data necessary for establishing final outcome were excluded from analysis. A separate version of each algorithm was trained for FL and CLL cases (Figure 6).

In the first iteration, a CNN was trained to predict the probability that a given patch represented a large-cell tumor rather than a small-cell tumor (Figure 7). These predictions were then fed to a k-nearest neighbor smoothing algorithm to encourage local homogeneity of patch predictions. In particular, each predicted probability was replaced by a weighted average of the original predicted value and the predicted values of surrounding patches (Figure 8). After this postprocessing step, the ratio of patches predicted to contain large-cell tumors to patches predicted to contain small-cell tumors was calculated for each slide. Finally, this feature was combined with the pathologist-estimated features mentioned above and input to an L1 regularized logistic regression algorithm trained to predict the patient’s probability of LCT.

In the second iteration, a CNN was trained to directly predict the probability that a given patch came from a patient who had undergone LCT, resulting in a complete end-to-end machine learning approach to the problem. As above, the resulting predictions were fed to a k-nearest neighbor smoothing algorithm to encourage local prediction homogeneity. For each patient, the mean predicted probability of LCT was calculated. As a final postprocessing step, the mean predicted patch probability for each patient in the training set (ie, 1 aggregate probability per patient) was used to train a “logit extremizer” corrector model originally developed to aggregate the knowledge of human predictors. The logit extremizer is a modified 1-parameter logistic regression model trained to predict patient LCT as a function of the mean predicted probability that a given patch came from a patient who had undergone LCT, resulting in a complete end-to-end machine learning approach to the problem. As above, the model corrects the effects of making decisions when there is imperfect information. In particular, the model assumes that if each predictor (patch) had access to the full set of information (the WSI), it would likely report a higher confidence. The model corrects this tendency using only 1 parameter α, improving model performance while posing a low risk of overfitting.9 In the model evaluation phase, the value of α that led to the smallest log loss (evaluated by comparing the logit extremizer outputs for a given value of α to the true patient LCT probabilities) across the training set was used to obtain a corrected probability for each patient in the held-out set. The corrected probability obtained after processing with the logit

Figure 4. A visualization of the stride augmentation procedure for patch count augmentation. Although here we demonstrate a reduced horizontal stride, we could equally well have reduced the vertical stride to create even more semioverlapping patches.

Figure 5. A visualization of the patch validity control process for white space detection.
extremizer model was treated as the final predicted LCT probability.

**Model Training and Architecture Specification**

In order to ensure generalizability of results, all models were trained on a patient-by-patient basis. In particular, a “leave one subject out” cross-validation testing scheme, in which a separate model was trained for each patient, with all patches relating to said patient removed from the training set, was used for all to make patch-level predictions. In this way, no patch prediction, and consequently no slide-level prediction, was based on a model trained on other patches from the same patient.

The 2 models each employed a similar network architecture, with either a ResNet-50 or MobileNet base model pretrained on ImageNet followed by 2 hidden layers of 1024 and 256 neurons in size, respectively. Unlike in a traditional transfer learning approach, the weights of the base CNN model were unfrozen so that the model could correct for the difference in feature distribution between the natural scene images typical of ImageNet and the medical slide images analyzed here. Batch normalization and dropout were used to improve model training and combat overfitting.

Both of the CNN-based algorithms were coded in a mix of Keras and Tensorflow and trained on a pair of NVIDIA GeForce RTX 2080 GPUs. Additional data augmentation was applied at train time, including random flips and crops. Each model was trained for 6 epochs of 200 batches each, where the batch size was set at 64. The patches for each batch were randomly sampled from the training set by using the following sampling scheme. First, the desired class label of each patch was selected by sampling from a binomial distribution with n equal to the batch size and a P value of 0.5. Then we randomly selected a patient whose slide contains images of the

---

**Figure 6.** Tumor heatmap of annotated whole slide images with (A) and without (B) large cell transformation. On the heatmap scale, red and blue represent large and small lymphoma cells, respectively.

**Figure 7.** ResNet-50–produced patches from annotated tumor regions displaying machine-generated quantification of large lymphoma cells. Large-cell probabilities were 0.001 (A), 0.498 (B), and 0.991 (C) (hematoxylin-eosin, original magnification ×500).
desired class and randomly selected a patch of the desired class from each patient. This sampling technique ensures that all patients are likely to be represented in each batch and that each batch contains an approximately equal number of patches of each class, which corrects for class imbalance issues in the original training data set. A stochastic gradient descent with restarts–based learning rate schedule was employed to prioritize locally convex local optima. At test time, an augmentation scheme was used in which each image was fed through the model 3 times with different random modifications. The results for each patch were then averaged to improve generalizability.

**Statistical Analysis and Prediction**

Statistical analysis of histologic outcomes was performed on features generated from pathologist evaluation of the slides and the output from the first iteration of the CNN model. L1 regularized logistic regression models were fit to determine which histologic variables correlated with patient outcome.

Three logistic regression models were then fit with the selected variables and compared, the first using pathologist estimates of large-cell tumor ratio, the second using the ratio of annotation surface area extracted from QuPath, and the third using the CNN’s predicted large-cell tumor ratio, in addition to the significant histologic variables. The models were used to calculate the odds ratio, P values, and an area under the receiver-operating characteristic (AUROC) curve. For clinical data, Fischer exact test and Mann-Whitney U tests (0.05; 2-tailed) were used to evaluate categoric and continuous (univariate) variables, respectively. Differences were claimed as statistically significant when the P value was less than .05.

Finally, the AUROC curve obtained from these 3 logistic regression models was compared to the curve calculated from the second CNN-based algorithm’s predictions of LCT probability to determine which of the 4 models was most effective in predicting LCT outcome.

**RESULTS**

**Follicular Lymphoma**

**Prediction of LCT and Histologic Scoring.**—In FL, bone marrow microscopy correlated with clinical transformation in 86% of cases (18 of 21); among these, LCT was observed in 67% (12 of 18). In 1 case, complete clinical data, including chemotherapy regimen, were unavailable. In the 2 discordant cases, bone marrow microscopy showed evidence of LCT in 1 case and no evidence of LCT in the other case. For the former patient, favorable clinical parameters were deemed to be most consistent with an indolent lymphoma, irrespective of the bone marrow findings. These parameters included a lack of constitutional symptoms, no lymphadenopathy or organomegaly on physical examination, normal LDH levels, and skeletal lesions with low standardized uptake values, not associated with significant pain. In the case with no evidence of LCT on bone marrow biopsy, an immediately subsequent lymph node biopsy was diagnostic of DLBCL. In cases with a concurrent lymph node or soft tissue biopsy, 80% of cases (4 of 5) were concordant. In the FL cohort altogether, 13 cases showed LCT, 7 cases did not, and 1 case was indeterminate because of unavailable clinical data.

Of the 8 histologic features evaluated on bone marrow biopsy, 4 were significantly associated with clinical transformation outcomes: proportion of large cells in the lymphoma (P = .02), chromatin pattern (P = .01), distinct nucleoli (P = .006), and proliferation index by Ki-67 (P = .04; Figure 9). Proportion of large cells in the lymphoma was most predictive. In cases with concordant bone marrow biopsy and clinical outcome data, all cases with LCT demonstrated 10% to 100% large lymphoma cells, open chromatin, distinct nucleoli, and Ki-67 indices from 10% to 85%. On the other hand, all cases without LCT demonstrated ≤10% large lymphoma cells and Ki-67 indices ≤5%. A cutoff of 10% large cells in FL has a sensitivity of 92% and a specificity of 88% in diagnosing LCT. Furthermore, 67% of cases (4 of 6) without LCT demonstrated condensed chromatin and nondistinct nucleoli. Histologic features that were not predictive included percent of marrow involved by lymphoma, irregular lymphoid aggregates, the presence of highly atypical cells, and fibrosis (Table 1).

**Clinical Data.**—When evaluating FL patients with clinical suspicion for LCT, those with evidence for LCT were noted to be older than those without evidence of LCT on biopsy (median age, 60 versus 54 years, P = .04). Furthermore, FL patients with evidence of transformation were found to have a higher rate of B symptoms (55% versus 0%, P = .04), and an increased LDH (513 versus 210, P = .30), as well as greater Ann Arbor stage (P = .002) and median National Comprehensive Cancer Network–IPI score

**Table 1. Summary of Scored Histologic Features in Cases of Follicular Lymphoma With Concordant Bone Marrow Biopsy and Clinical Data**

| Feature                        | LCT       | No LCT     | P Value |
|-------------------------------|-----------|------------|---------|
| Marrow involvement by lymphoma, % | 15–100    | 10–55      | .30     |
| Proportion of large lymphoma cells, % | 10–100    | ≤2 to 10   | .02*    |
| Ki-67, %                      | 10–85     | ≤5         | .04*    |
| Highly atypical cells         | 5/12      | 1/6        | >.99    |
| Fibrosis (1–3)                |           |            | .99     |
| 1                             | 4/12      | 4/6        |        |
| 2                             | 6/12      | 1/6        |        |
| 3                             | 1/12      | 1/6        |        |
| Not available                 | 1/12      |            |        |
| Irregular lymphoid aggregates | 6/12      | 0/6        | .20     |
| Distinct nucleoli             | 12/12     | 1/6        | .006*   |
| Open chromatin                | 12/12     | 1/6        | .01*    |

**Abbreviation:** LCT, large cell transformation.

* Statistical significance.

**Predictions of Lymphoma Transformation**—Irshaid et al

---

**Figure 8. Visualization of the k-nearest neighbor (kNN) smoothing process.** An isolated patch of predicted class large-cell tumor (blue) is converted via kNN smoothing to the class of its nearest neighbors.
at the time of LCT. There was a trend toward a higher IPI score for those patients with evidence of LCT \((P = .05; \text{Table 2})\). The LDH at FL diagnosis was not predictive of eventual LCT \((P = .41)\).

**Machine Learning/Predictive Model Comparison.**—Of the 4 models considered, the end-to-end CNN-based model obtained the best results, with an AUROC of 0.857. This was followed by the logistic regression model trained on surface area estimates extracted from QuPath annotations (AUROC, 0.851). The model trained on CNN-estimated large-cell ratio (AUROC, 0.80) and the model trained only on pathologist-estimated features (AUROC, 0.839) fared somewhat worse (Figure 10). Combining the CNN models with additional pathologist features resulted in improved performance, with an AUROC of 0.923 for the end-to-end model and an AUROC of 0.881 for the CNN-estimated large-cell ratios. Combining the other models with additional pathologist features resulted in more modest performance gains, with an AUROC of 0.851 for the pathologist-only model and an AUROC of 0.857 for the QuPath-assisted model. Pathologist and machine learning–derived estimates of proportion of large cells per case had a correlation of 0.64. Compared with pathologist-derived estimates, machine-generated quantification of proportion of large cells per case displayed better reproducibility and a stronger correlation with final clinical outcome data. The reproducibility is likely stronger because of

**Table 2. Comparisons of Patient- and Disease-Specific Metrics Between Cases With and Without Large-Cell Transformation (LCT) at the Time of Follicular Lymphoma Diagnosis and at the Time of Clinical Suspicion of Transformation**

| Clinical Metrics | Cases With LCT \((n = 13)\) | Cases Without LCT \((n = 7)\) | \(P\) Value |
|------------------|-----------------------------|-------------------------------|-----------|
| **At follicular lymphoma diagnosis** | | | |
| Median age, y    | 47 (range, 34–71)           | 53 (range, 31–77)             | .96       |
| Median FLIPI score | 3 (range, 2–4)              | 5 (range, 2–5)                | .39       |
| Median grade     | 2                           | 2                             | N/A       |
| Median Ann Arbor stage | 4                  | 4                             | N/A       |
| Median LDH       | 201                         | 171                           | .41       |
| **At clinical suspicion of LCT** | | | |
| Median age, y    | 54 (range, 42–71)           | 60 (range, 55–79)             | .04*      |
| Time since diagnosis, mo | 20.6               | 197.2                        | .43       |
| B-symptoms, %    | 55                          | 55                            | .04*      |
| LDH              | 513                         | 210                           | .30       |
| Increasing LDH level, % | 55                     | 0                             | .04*      |
| Median Ann Arbor stage | 4 (range, 4–4)       | 3 (range, 2–4)               | .002*     |
| Median IPI score | 2 (range, 1–4)             | 1 (range, 1–2)               | .05*      |
| Median NCCN-IPI score | 4 (range, 3–7)     | 2 (range, 2–4)               | .02*      |

Abbreviations: FLIPI, Follicular Lymphoma International Prognostic Index; IPI, International Prognostic Index; LDH, lactate dehydrogenase; N/A, not applicable; NCCN-IPI, National Comprehensive Cancer Network–IPI.

* Statistical significance.
the fact that the model uses its learned pattern in the same fashion, compared with a pathologist who may place more weight on certain aspects when evaluating certain cases, or even evaluating the same case on different days.

**Chronic Lymphocytic Leukemia**

**Prediction of LCT and Histologic Scoring.**—In CLL, bone marrow microscopy correlated with clinical transformation in 72% of cases (36 of 50) and was discordant in 10% of cases (5 of 50); among the concordant cases, LCT was observed in 22% (8 of 36). Final clinical transformation data were not available for the remaining 18% of cases (9 of 50). In the 5 discordant cases, bone marrow microscopy showed evidence of LCT in 1. Aggressive chemotherapy was not administered in this case because of favorable clinical parameters and spontaneous resolution of lymphadenopathy. In the 4 discordant cases with no evidence of LCT on bone marrow microscopy, aggressive chemotherapy was administered for various reasons: 2 cases had concurrent lymph node biopsies with diagnostic evidence of DLBCL, 1 case had a history of DLBCL, and 1 case displayed resistance to first-line therapy, necessitating progression to other regimens. In cases with a concurrent lymph node or soft tissue biopsy, 83% of cases (10 of 12) were concordant. In the CLL cohort altogether, 12 cases showed LCT, 29 cases did not, and 9 cases were indeterminate because of unavailable clinical data.

Of the 8 histologic features tested, 5 were significantly associated with clinical LCT outcomes: distinct nucleoli (P = .003), pathologist estimates of the proportion of large lymphoma cells (P = .01), Ki-67 proliferation index (P = .03), presence of highly atypical cells (P = .003), and chromatin pattern (P = .006; Figure 11). In cases with concordant bone marrow biopsy findings and clinical outcome data, cases with LCT demonstrated 30% to 100% large lymphoma cells and Ki-67 indices of 5% to 90%; in contrast, cases without LCT demonstrated ≤5% large lymphoma cells and Ki-67 indices of ≤20%. A cutoff of 30% large cells in CLL has a sensitivity of 100% and a specificity of 85% in diagnosing LCT. Highly atypical cells were noted in 88% of cases (7 of 8) with LCT but were not noted in any of the cases without LCT, making it a moderately sensitive but highly specific feature of LCT. Similarly, 88% of cases (7 of 8) with LCT demonstrated open chromatin and distinct nucleoli; in contrast, all cases without LCT demonstrated condensed chromatin and nondistinct nucleoli. Histologic features that were not predictive included percent of marrow core involved by lymphoma, irregular lymphoid aggregates, and fibrosis (Table 3).

**Clinical Data.**—When evaluating only CLL patients, those with evidence of Richter transformation were noted to have a higher rate of del(17p) at initial diagnosis (54% versus 13%, P = .02). Rai stage, β2-microglobulin, CD38 and IGHV status, and LDH at diagnosis were not predictive of eventual Richter transformation (Table 4). Of those CLL patients with clinical suspicion for Richter transformation, those with evidence of such on biopsy were noted to have a higher median Ann Arbor state (4 versus 3, P < .001) and median National Comprehensive Cancer Network–IPI score (4 versus 3, P = .02). There was a trend toward a higher LDH and increasing LDH at suspicion of transformation for those with morphologic evidence in comparison with those without such evidence, but these were not statistically significant findings (Table 4).

**Machine Learning/Predictive Model Comparison.**—Of the 4 models considered, when in isolation, the QuPath-estimated large-cell ratio model proved the most predictive (AUROC, 0.815) for CLL transformation by final clinical determination. This was followed by the end-to-end CNN model and the pathologist-only model, both of which had an AUROC of 0.733. The first iteration of the CNN model performed the worst, with an AUROC of 0.720. However, combining the predictions of any of these models with other histologic features resulted in better performance for both deep learning models and the pathologist-only model, with

---

**Table 3. Summary of Scored Histologic Features in Cases of Chronic Lymphocytic Leukemia With Concordant Bone Marrow Biopsy and Clinical Data**

| Feature                        | LCT       | No LCT    | P Value |
|--------------------------------|-----------|-----------|---------|
| Marrow involvement by lymphoma, % | 15–100    | 5–100     | .09     |
| Proportion of large lymphoma cells, % | 30–100    | ≤5        | .01*    |
| Ki-67, %                        | 5–90      | ≤20       | .03*    |
| Highly atypical cells           | 7/8       | 0/28      | .003*   |
| Fibrosis (1–3)                  |           |           | .99     |
| 1                              | 5/8       | 20/28     |         |
| 2                              | 2/8       | 5/28      |         |
| 3                              | 1/8       | 3/28      |         |
| Irregular lymphoid aggregates   | 4/8       | 13/28     | .79     |
| Distinct nucleoli              | 7/8       | 0/28      | .003*   |
| Open chromatin                 | 7/8       | 0/28      | .006*   |

Abbreviation: LCT, large cell transformation. *Statistical significance.
Figure 11. Bone marrow pathology of chronic lymphocytic leukemia suspicious for transformation. Bone marrow core biopsies demonstrating irregular lymphoid aggregates (A), cells with marked nuclear atypia and pleomorphism (B), and large lymphoma cells with open chromatin and distinct nucleoli (C) (hematoxylin-eosin, original magnifications ×40 [A] and ×400 [B and C]).

| Clinical Metrics                          | Cases With LCT (n = 12) | Cases Without LCT (n = 29) | P Value* |
|-------------------------------------------|-------------------------|----------------------------|----------|
| At chronic lymphocytic leukemia diagnosis |                         |                            |          |
| Median age, y                             | 59 (range, 44–80)       | 65 (range, 45–87)          | .42      |
| Rai stage                                 | 2.5                     | 2                          | .84      |
| β2-microglobulin >3.5, %                  | 40                      | 75                         | .29      |
| CD38 <30%, %                              | 63                      | 73                         | .39      |
| IGHV mutated, %                           | 0                       | 0                          | N/A      |
| Complex karyotype, %                      | 0                       | 10                         | .52      |
| Del(17p) present, %                       | 54                      | 13                         | .02*     |
| Median LDH                                | 231                     | 234                        | .42      |
| At clinical suspicion of LCT              |                         |                            |          |
| Median Age, y                             | 69 (range, 52–80)       | 65 (range, 50–88)          | .98      |
| Time since diagnosis, mo                  | 31.6                    | 46.2                       | .71      |
| B-symptoms, %                             | 33                      | 29                         | .74      |
| LDH                                       | 319                     | 287                        | .06      |
| Increasing LDH level, %                   | 56                      | 20                         | .10      |
| Median Ann Arbor stage                    | 4 (range, 3–4)          | 3 (range, 2–4)             | <.001*   |
| Median IPI score                          | 2.5 (range, 1–4)        | 2 (range, 1–4)             | .20      |
| Median NCCN-IPI score                     | 4 (range, 3–8)          | 3 (range, 2–7)             | .02*     |

Abbreviations: IGHV, immunoglobulin heavy chain variable region; IPI, International Prognostic Index; LDH, lactate dehydrogenase; N/A, not applicable; NCCN-IPI, National Comprehensive Cancer Network-IPI.

* Statistical significance.
predictions of lymphoma transformation from CLL or low-grade FL. Histologic parameters of FL that showed significant association with LCT included the presence of distinct nucleoli, percent of large cells, open chromatin, and proliferation index. Histologic features of CLL that showed significant association with LCT were the same and included 1 additional characteristic in the presence of highly atypical cells. Although this only occurred in a few cases, it was a highly specific finding. Although these findings may be intuitive for the practicing pathologist, it can be helpful to break the specific elements that contribute to this distinction via light microscopy. Although this is a single-institution study and the number of cases is limited, our findings suggest a cutoff of 10% large cells to diagnose LCT in FL and 30% in CLL. However, a larger validation set would further aid in establishing a cutoff of proportion of large cells or proliferation index needed in the diagnosis of LCT on marrow specimens.

The proportion of large cells in the biopsy is of critical importance to the diagnosis of LCT in the context of any low-grade B-cell lymphoma. It is interesting to note that the presence of distinct nucleoli and open chromatin by pathologist examination were also consistently predictive across the 2 lymphoma subtypes. This can be due to the relatively more difficult task of quantifying large-cell percentage in overall tumor population. However, given the overall reasonable interrater reliability with regard to nuclear features in our study, it may also be that they are indeed intrinsically associated with transformation. For instance, prolymphocytes are often distinguished by their open chromatin and identifiable nucleoli. Interestingly, the finding of highly atypical cells also appeared to be highly predictive of LCT in CLL biopsies. From the morphologist’s perspective, this might be helpful in that most CLLS are composed of uniformly round cells at baseline. Hence, in the rare cases where highly irregular nuclear contours are seen, it is histologically striking and may be compatible with the cytology of paraimmunoblasts. It has been reported that in patients with Richter transformation, atypical lymphocyte morphology is often seen in blood or marrow, even when the biopsy is not diagnostic.7 This particular feature was not assessed in FL because indolent FL is at baseline, composed of cells with heterogeneous nuclear contours.

Most lymphoma subclassification requires lymph node or soft tissue biopsy, which may be followed by bone marrow staging biopsy. For instance, the diagnosis of splenic marginal zone lymphoma or splenic diffuse red pulp small B-cell lymphoma rests on the histology seen in splenectomy samples. However, recent studies show that bone marrow histology can provide substantial evidence for confirmation of these diagnoses even though the spleen remains the gold standard.10 In older literature, marrow histologic patterns were the best prognostic parameter in indolent neoplasms, such as CLL.11

Although transformation of low-grade lymphoma to large B-cell lymphoma is not pathologically defined in the context of bone marrow biopsy, the finding of discrepant morphology (ie, sheets of large cells in bone marrow versus small cells in tissue biopsy) has been reported.13 Hence, the distinction of concordant versus discordant marrow infiltration by DLBCL is considered possible by histology alone.13,14 This evaluation is based largely on the finding of either mostly large, noncleaved cells versus mostly small, mature-appearing lymphoma cells.15 Hence, it should also be possible to see discordant/transformed marrow morphology in the context of a patient with known low-grade lymphoma. Furthermore, primary bone marrow DLBCL has been well described. It has been suggested that some of these primary bone lymphomas may have arisen from nodal or tissue-based indolent lymphomas, such as FL or CLL.16

**DISCUSSION**

This is the first study to systematically assess the various components of marrow evaluation for large-cell lymphoma transformation, as well as the first to provide a computational model for this pathologic diagnosis. Given that the pathologic diagnosis of LCT is rarely straightforward outside the context of large excisional lymph node biopsies, we sought to (1) parse out the elements that could best inform these decisions, (2) demonstrate how deep learning algorithms using WSI can contribute, and (3) compare how each performs in predicting a clinical event as well as in collaboration.

We scored each histologic finding on the biopsy for suspicion of transformation from CLL or low-grade FL. A top AUROC of 0.823 for the end-to-end CNN model and pathologist-only models, and an AUROC of 0.780 for the QuPath-estimated large-cell ratio CNN model. The QuPath model with the addition of histologic features achieved a performance similar to the model trained in isolation (Figure 12). Pathologist- and machine-derived estimates of the proportion of large cells per case had a correlation of 0.765. Although certain cases were predicted to have undergone LCT through the QuPath-estimated large-cell ratio, although not by initial traditional methods, review of these cases did not reveal additional features by light microscopy that could be differentiated from nontransformed cases via traditional pathologic assessment.
Tissue confirmation of transformation via lymph node biopsy can sometimes be confounded by the biopsy of nonlesional adjacent tissue or a partial biopsy that is difficult to interpret. Because there are occasionally urgent clinical situations that render tissue biopsy difficult to obtain, pathologists have been asked to assess for transformation on bone marrow biopsy. A caveat should be made that absence of evidence in marrow should not rule out the possibility of transformation that could be demonstrated by gold standard lymph node biopsy.

Early studies using computer-assisted systems to evaluate lymphoma images demonstrated that color-texture analysis approaches that combine model-based intermediate representation of cytology can distinguish the most aggressive FL (grade 3 of 3) from its low-grade counterparts with high sensitivity and specificity, and can classify FL, CLL and mantle cell lymphoma. Furthermore, a large histopathology-based study to recognize FL versus reactive hyperplasia showed that machine learning techniques were reliable in screening for FL, even those that were BCL2. These tools were promising but used pathology criteria alone as the gold standard for comparison.

A strength of this study is that the final clinical determination of lymphoma transformation was the gold standard, rather than biopsy alone. Although the clinician’s decision on transformation takes biopsy results into consideration, it also includes many clinical and imaging parameters that finally inform the decision to treat with multiagent chemotherapy. More recent experiments carried out in solid-tumor malignancies have demonstrated the power of machine learning methods in predicting prognosis based on tissue images. We have found that combining the CNN models with pathologist-interpreted features resulted in the best prediction of whether a patient with indolent lymphoma (FL or CLL) underwent disease progression. If larger data sets validate these initial findings, a consensus proposal may suggest that an end-to-end CNN model complement more easily established pathologic features, such as an elevated percent of large cells (>10% in FL and >30% in CLL) and elevated proliferation index (>10% in FL and >20% in CLL). Furthermore, the machine-assisted models show greater reproducibility over time. With challenging cases or cases in settings where subspecialty consultation may not be obtainable in a timely manner, these novel tools may be an important aid.

A new study in the field of dermatologic lesions found that good-quality artificial intelligence support of image-based clinical diagnoses enhanced the accuracy of physicians, particularly those with less expertise, similarly to obtaining a second opinion. In addition, they found that aggregated artificial intelligence–based multiclass probabilities increased accuracy compared with individual raters or artificial intelligence alone. Similarly, deep learning algorithms have been shown to demonstrate high sensitivity and specificity in detecting diabetic retinopathy and macular edema in retinal fundus photographs. Given the explosion of telemedicine use due to the recent coronavirus disease 2019 (COVID-19) pandemic along with a concomitant shortage of available physicians, machine learning support for diagnostic work has become even more desirable.

Although the large size of WSIs presents significant challenges for direct analysis of histopathologic images by traditional machine learning image processing methods, several studies have managed to adapt the input image data to make such analysis possible. State-of-the-art WSI analysis algorithms rely largely on adapting the input WSI data into a format that allows one to leverage the flexibility and predictive power of CNNs, widely used in other image analysis tasks. Such a CNN-based algorithm achieved nearly perfect results on the CAMELYON16 competition data set of lymph node metastases. Although more work is needed to adapt deep learning algorithms to the unique challenges posed by WSIs, a recent paper that proposes the use of “steerable filters” to exploit rotational symmetry in WSIs is a step in the right direction.

Our results are intriguing not only in demonstrating the proficiency of machine learning algorithms in its application to lymphoma diagnoses and prognosis, but also in highlighting the potential for differences in performance across tumor types. In a recent study using a large data set of non–small-cell lung cancer tissues, fully automated machine learning methods were capable of classification without relying on prior pathology knowledge. Nevertheless, the authors found that the detection of squamous cell carcinoma appeared to be superior to detection of other subtypes, because of the relative homogeneity of the cohort. Likewise, it is possible that we found better predictive machine learning results with FL than with CLL because of aspects of histology in FL that are more conserved across patients.

Although novel and emerging AI-based algorithms may not be available for clinically validated use in our immediate future, these findings support the need for further studies that integrate traditional histopathologic examination with CNN. Machine learning can make the biggest impact in fields where expertise is limited, data are abundant, and final predictions are actionable. A recent study demonstrates the utility of using machine learning to capture information from gene expression analysis of B-cell lymphomas to aid in the subclassification by hemopathology. Such assays are likely to become diagnostic companion tools and aid in the identification of targetable pathways. Potential hurdles for validating CNN use in a field with abundant subtypes is that large training sets have been shown to work best. Multi-institutional collaborations to annotate rare tumors would be necessary, such as demonstrated in a study of computational analysis in TFE3 translocated renal cell carcinoma.

Because clinical decisions rely, at least in part, on pathologic findings, the biopsy results often impact the ultimate decision of whether that patient’s disease has progressed and whether treatment must be initiated. This impact may confound the study of correlation between large-cell morphology and clinical outcome. However, we note that because of the lack of clear criteria, very few of the biopsy reports for either FL or CLL actually come to a definitive diagnosis of transformation. Hence, aside from lymph node or soft tissue biopsy adequate for that diagnosis, clinicians typically have substantial leeway to compile clinical, laboratory, and radiologic evidence in order to come to an overarching conclusion.

Finally, the training that is essential to creating a helpful deep learning algorithm is very time-consuming and requires a substantial collection of well-annotated samples. This may prove to be a large hurdle for such artificial intelligence to become relevant in the diagnoses of rare and increasingly subdivided categories of disease.

We thank Fahad Ahmed, MD, in the laboratory of David Rimm, MD, PhD, for assistance in image processing, and Yale Pathology Tissue Services for assistance in archival specimen retrieval.
with follicular lymphoma who had histologic transformation after response to
treatment.

2575–2582.

bone marrow in patients with chronic lymphocytic leukemia.

Haematologica outcomes of Richter transformation: experience of 204 patients from a single
center. 2020;105(3):765–773.

The clinical significance of large cells in bone marrow in patients with chronic lymphocytic leukemia. Cancer. 2004; 100(10):2167–2175.

combining multiple probability predictions using a simple logit model. Int J Forecast. 2014;30(2):344–356.

Bone marrow histopathology in the diagnostic evaluation of splenic marginal-zone and splenic diffuse red pulp small
B-cell lymphoma: a reliable substitute for spleen histopathology? Am J Surg Pathol. 2012;36(11):1609–1618.

Bone marrow histologic pattern—the best single prognostic parameter in chronic lymphocytic leukemia: a multivariate survival analysis of 329 cases. Blood. 1984;64(3):642–648.

Primary splenic low-grade follicular lymphoma presenting with leukoemia and large cell transformation in the marrow. Pathology. 2017;49(6):649-652.

Concordant but not discordant bone marrow involvement in diffuse large B-cell lymphoma predicts a poor clinical outcome independent of the International Prognostic Index. Blood. 2007;110(4):1278-1282.

Impact of concordant and discordant bone marrow involvement on outcome in diffuse large B-cell lymphoma treated with R-CHOP. J Clin Oncol. 2011;29(11):1452–1457.

Concordant bone marrow involvement of diffuse large B-cell lymphoma represents a distinct clinical and biological entity in the era of immunotherapy. Leukemia. 2018;32(2):353–363.

Primary bone marrow diffuse large B cell lymphoma: a case series and review. Ann Hematol. 2011;90(7):791–796.

Clinical characteristics and outcomes of Richter transformation: experience of 204 patients from a single center. Haematologica. 2020;105(3):765–773.

The clinical significance of large cells in bone marrow in patients with chronic lymphocytic leukemia. Cancer. 2004; 100(10):2167-2175.

Arch Pathol Lab Med Predictions of lymphoma Transformation—Irshaid et al

References

1. Robboy SJ, Weintraub S, Horwath AE, et al. Pathologist workforce in the United States, I: development of a predictive model to examine factors influencing supply. Arch Pathol Lab Med. 2013;137(12):1723–1732.

2. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2015. CA Cancer J Clin. 2015;65(1):5–29.

3. Dreyling M, Ghielmini M, Rule S, Salles G, Vitolo U, Ladetto M. Newly

diagnosed and relapsed follicular lymphoma: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol. 2016;27(suppl 5):83–90.

4. Sarkozy C, Tremeny M, Xerri L, et al. Risk factors and outcomes for patients with follicular lymphoma who had histologic transformation after response to
first-line immunochemotherapy in the PRIMA trial. J Clin Oncol. 2016;34(22): 2575–2582.

5. Hallek M. Chronic lymphocytic leukemia: 2020 update on diagnosis, risk
stratification and treatment. Am J Hematol. 2019;94(11):1266–1287.

6. Wang Y, Tschautscher MA, Rabe KG, et al. Clinical characteristics and outcomes of Richter transformation: experience of 204 patients from a single center. Haematologica. 2020;105(3):765–773.

7. Ma Y, Mansour A, Bekele BN, et al. The clinical significance of large cells in bone marrow in patients with chronic lymphocytic leukemia. Cancer. 2004; 100(10):2167-2175.

8. Bankhead P, Loughrey MB, Fernández J, et al. QuPath: open source software for digital pathology image analysis. Sci Rep. 2017;7(1):1–7.

9. Satopää VA, Baron J, Foster DP, Mellers BA, Teflock PE, Ungar LH. Combining multiple probability predictions using a simple logit model. Int J Forecast. 2014;30(2):344–356.

10. Ponzoni M, Kanelis G, Pouliou E, et al. Bone marrow histopathology in the diagnostic evaluation of splenic marginal-zone and splenic diffuse red pulp small
B-cell lymphoma: a reliable substitute for spleen histopathology? Am J Surg Pathol. 2012;36(11):1609–1618.

11. Rozman C, Montserrat E, Rodriguez-Fernandez JM, et al. Bone marrow histologic pattern—the best single prognostic parameter in chronic lymphocytic leukemia: a multivariate survival analysis of 329 cases. Blood. 1984;64(3):642–648.

12. Jhuang JY, Hsieh YC, Kuo CC, Su YZ, Chuang SS. Primary splenic low-grade follicular lymphoma presents with leukoemia and large cell transformation in the marrow. Pathology. 2017;49(6):649-652.

13. Chung R, Lai R, Wei P, et al. Concordant but not discordant bone marrow involvement in diffuse large B-cell lymphoma predicts a poor clinical outcome independent of the International Prognostic Index. Blood. 2007;110(4):1278-1282.

14. Sehn LH, Scott DW, Chhanabhai M, et al. Impact of concordant and discordant bone marrow involvement on outcome in diffuse large B-cell lymphoma treated with R-CHOP. J Clin Oncol. 2011;29(11):1452–1457.

15. Yao Z, Deng L, Xu-Monette ZY, et al. Concordant bone marrow involvement of diffuse large B-cell lymphoma represents a distinct clinical and biological entity in the era of immunotherapy. Leukemia. 2018;32(2):353–363.

16. Chang H, Hung YS, Lin TL, et al. Primary bone marrow diffuse large B cell lymphoma: a case series and review. Ann Hematol. 2011;90(7):791–796.

17. Sertel O, Kong J, Catalyurek UV, Lozanski G, Saltz JH, Gurcan MN. Histopathological image analysis using model-based intermediate representations and color texture: follicular lymphoma grading. J Signal Process Syst. 2009; 53(1–3):169.

18. Janowczyk A, Madabhushi A. Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases. J Pathol Inform. 2016; 7:29.

19. Syrykh C, Abreu A, Amara N, et al. Accurate diagnosis of lymphoma on whole-slide histopathology images using deep learning. NPJ Digit Med. 2020; 3(1):1-8.

20. Yu KH, Zhang C, Berry GJ, et al. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. Nat Commun. 2016;7(1):12474.

21. Tschandl P, Rinner C, Apalla Z, et al. Human-computer collaboration for skin cancer recognition. Nat Med. 2020;26(8):1229–1234.

22. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016;316(22):2402-2410.

23. Komura D, Ishikawa S. Machine learning methods for histopathological image analysis. Comput Struct Biotechnol J. 2018;6:34–42.

24. Bejnordi BE, Veta M, Van Diest PJ, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA. 2017;318(22):2199-2210.

25. Graham S, Epstein DJ, Rajpoot N. Dense steerable filter CNNs for exploiting rotational symmetry in histology images. arXiv. 2004.03037. Posted online on April 6, 2020.

26. Yu KH, Wang F, Berry GJ, et al. Classifying non-small cell lung cancer types and transcriptomic subtypes using convolutional neural networks. J Am Med Inform Assoc. 2020;27(5):757-769.

27. Bobée V, Drieuks F, Marchand V, et al. Combining gene expression profiling and machine learning to diagnose B-cell non-Hodgkin lymphoma. Blood Cancer J. 2020;10(5):59.

28. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553): 436–444.

29. Madabhushi A, Lee G. Image analysis and machine learning in digital pathology: challenges and opportunities. Med Image Anal. 2016;33:170-175.

30. Cheng J, Han Z, Mehra R, et al. Computational analysis of pathological images enables a better diagnosis of TFE3 Xp11.2 translocation renal cell carcinoma. Nat Commun. 2020;11(1):1778.