

ORIGINAL RESEARCH

Assessment of the current and emerging criteria for the histopathological classification of lung neuroendocrine tumours in the lungNENomics project

É. Mathian^{1,2}, Y. Drouet^{3,4}, A. Sexton-Oates¹, M. G. Papotti⁵, G. Pelosi⁶, J.-M. Vignaud^{7,8}, L. Bric⁹, A. Mansuet-Lupo¹⁰, F. Damiola¹¹, C. Altun¹¹, J.-P. Berthet¹², C. B. Fournier¹³, O. T. Brustugun^{14,15}, G. Centonze¹⁶, L. Chalabreysse¹⁷, V. T. de Montpréville¹⁸, C. M. di Micco¹⁹, E. Fadel^{18,20}, N. Gadot¹¹, P. Graziano¹⁹, P. Hofman²¹, V. Hofman²¹, S. Lacomme⁸, M. Lund-Iversen²², L. Mangiante^{1,23}, M. Milione¹⁶, L. A. Muscarella¹⁹, C. Perrin¹⁷, G. Planchard²⁴, H. Popper⁹, N. Rousseau¹³, L. Roz²⁵, G. Sabella¹⁶, S. Tabone-Eglinger²⁶, C. Voegelé¹, M. Volante⁵, T. Walter²⁷, A.-M. Dingemans²⁸, L. Moonen²⁹, E. J. Speel²⁹, J. Derks³⁰, N. Girard³¹, L. Chen^{2,32}, N. Alcalá¹, L. Fernandez-Cuesta^{1*†}, S. Lantuejoul^{11†} & M. Foll^{1†}

¹Rare Cancers Genomic Team, Genomic Epidemiology Branch, International Agency for Research on Cancer (IARC-WHO), Lyon, France; ²Department of Mathematics and Informatics, Ecole Centrale de Lyon, Lyon, France; ³UMR CNRS 5558 LBBE, Claude Bernard Lyon 1 University, Villeurbanne, France; ⁴Prevention & Public Health Department, Centre Léon Bérard, Lyon, France; ⁵Department of Oncology, University of Turin, Turin, Italy; ⁶Department of Oncology and Hemato-Oncology, University of Milan, Milan, Italy; ⁷Department of Biopathology, Institut De Cancérologie de Lorraine (CHRU-ICL), Vandoeuvre-lès-Nancy, France; ⁸University Hospital of Nancy (CHRU), Nancy, France; ⁹Diagnostic and Research Institute of Pathology, Medical University of Graz, Graz, Austria; ¹⁰Department of Pathology, Hôpital Cochin, AP-HP, Université de Paris, Paris, France; ¹¹Department of Biopathology, Centre Léon Bérard & Pathology Research Platform, Cancer Research Center of Lyon, Lyon, France; ¹²Department of Thoracic Surgery, FHU OncoAge, Nice Pasteur Hospital, University Cote d'Azur, Nice, France; ¹³Caen Lower Normandy Tumour Bank, Centre François Baclesse, Caen, France; ¹⁴Section of Oncology, Drammen Hospital, Vestre Viken Hospital Trust, Drammen, Norway; ¹⁵Institute of Clinical Medicine, University of Oslo, Oslo, Norway; ¹⁶First Pathology Division, Department of Pathology and Laboratory Medicine, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy; ¹⁷Hospices Civils de Lyon, GHE, Institut de Pathologie Est, Bron, France; ¹⁸Department of Pathology, Hôpital Marie-Lannelongue, Groupe Hospitalier Paris Saint Joseph, Le Plessis Robinson, France; ¹⁹Unit of Oncology, Fondazione IRCCS Cas Sollevio della Sofferenza, San Giovanni Rotondo, Italy; ²⁰Department of Thoracic and Vascular Surgery and Heart-Lung Transplantation, Université Paris-Saclay, Le Plessis-Robinson, France; ²¹FHU OncoAge, Biobank BB-0033-0025, Laboratory of Clinical and Experimental Pathology, Nice Pasteur Hospital, University Cote d'Azur, Nice, France; ²²Department of Pathology, Oslo University Hospital, Oslo, Norway; ²³School of Medicine, Stanford University, Stanford, USA; ²⁴Pathology Department, Caen University Hospital, Normandy University, Caen, France; ²⁵Tumour Genomics Unit, Department of Research, Fondazione IRCCS Istituto Nazionale Dei Tumori, Milan, Italy; ²⁶Biological Resource Center, Centre Léon Bérard, Lyon, France; ²⁷Service d'Oncologie Médicale, Groupement Hospitalier Centre, Institut de Cancérologie des Hospices Civils de Lyon, Lyon, France; ²⁸Department of Pulmonary Medicine, Erasmus MC Cancer Institute, University Medical Center, Rotterdam, The Netherlands; ²⁹Department of Pathology, GROW School for Oncology and Reproduction, Maastricht University Medical Centre, Maastricht, Netherlands; ³⁰Department of Pulmonary Diseases, GROW School for Oncology and Reproduction, Maastricht University Medical Centre, Maastricht, The Netherlands; ³¹Institut Curie, Versailles, France; ³²Institut Universitaire de France (IUF), Paris, France



Available online xxx

Background: Six thoracic pathologists reviewed 259 lung neuroendocrine tumours (LNETs) from the lungNENomics project, with 171 of them having associated survival data. This cohort presents a unique opportunity to assess the strengths and limitations of current World Health Organization (WHO) classification criteria and to evaluate the utility of emerging markers.

Patients and methods: Patients were diagnosed based on the 2021 WHO criteria, with atypical carcinoids (ACs) defined by the presence of focal necrosis and/or 2-10 mitoses per 2 mm². We investigated two markers of tumour proliferation: the Ki-67 index and phospho-histone H3 (PHH3) protein expression, quantified by pathologists and automatically via deep learning. Additionally, an unsupervised deep learning algorithm was trained to uncover previously unnoticed morphological features with diagnostic value.

Results: The accuracy in distinguishing typical from ACs is hampered by interobserver variability in mitotic counting and the limitations of morphological criteria in identifying aggressive cases. Our study reveals that different Ki-67 cut-offs can categorise LNETs similarly to current WHO criteria. Counting mitoses in PHH3+ areas does not improve diagnosis, while providing a similar prognostic value to the current criteria. With the advantage of being time efficient, automated assessment of these markers leads to similar conclusions. Lastly, state-of-the-art deep learning modelling does not uncover undisclosed morphological features with diagnostic value.

*Correspondence to: Dr Lynnette Fernandez-Cuesta, 25 Avenue Tony Garnier, Lyon 69007, France. Tel: +33-6-95-29-91-23
E-mail: fernandezcuestal@iarc.who.int (L. Fernandez-Cuesta).

†These authors contributed equally to this work.

2059-7029/© 2024 World Health Organization; licensee Elsevier Ltd. This is an open access article under the CC BY-NC-ND IGO license (<http://creativecommons.org/licenses/by-nc-nd/3.0/igo/>).

Conclusions: This study suggests that the mitotic criteria can be complemented by manual or automated assessment of Ki-67 or PHH3 protein expression, but these markers do not significantly improve the prognostic value of the current classification, as the AC group remains highly unspecific for aggressive cases. Therefore, we may have exhausted the potential of morphological features in classifying and prognosticating LNETs. Our study suggests that it might be time to shift the research focus towards investigating molecular markers that could contribute to a more clinically relevant morpho-molecular classification.

Key words: lung neuroendocrine tumours, histological classification, deep learning, Ki-67, PHH3

INTRODUCTION

Lung neuroendocrine neoplasms (NENs) are divided, according to the World Health Organization (WHO),¹ into high-grade neuroendocrine carcinomas (NEC), which encompass small-cell lung carcinoma and large-cell neuroendocrine carcinomas, and well-differentiated neuroendocrine tumours (NETs). These morphological types respectively account for 15%, 3%, and 2% of all lung cancer cases. LNETs can be further subdivided into low-grade (G1) NET or typical carcinoids (TCs) and intermediate-grade (G2) NET or atypical carcinoids (ACs). ACs are more aggressive than TCs, with a fourfold to sixfold higher risk of metastatic disease and relapse within 10 years after surgery. According to the WHO classification, the number of mitoses and presence of necrosis are the criteria to distinguish G1 (TC) from G2 (AC) LNETs. TCs have <2 mitoses per 2 mm², whereas ACs have 2-10 mitotic figures. Any focus of necrosis is diagnostic of AC. The relevance of a classification system lies in its ability to discriminate the most aggressive cases specifically and sensitively by classifying them as ACs, so that these patients can be monitored for longer. Incorrect classification may result in costly, distressing, and unnecessary prolonged clinical management, which is mainly explained by the low reproducibility of mitotic count and by the challenge to assess necrosis in a reproducible manner.^{2,3}

To address this problem, the well-known Ki-67 proliferation index has been tested as a marker to improve reproducibility.⁴ The 2022 WHO Classification of Endocrine Tumours introduced Ki-67 as a pillar for NETs of all body sites. In the specific case of the lung, the use of Ki-67 is encouraged but not mandatory, even if considered useful in distinguishing highly proliferative ACs.^{1,5-7}

Ki-67 is expressed during G1, S, G2, and M phases of the cell cycle, so its correlation with mitotic count is not perfect.⁸ Phospho-histone H3 (PHH3) expression, on the other hand, is restricted only to the M phase.⁹ This marker, which has already been extensively studied for the classification of gastrointestinal and pancreatic NETs, allows sensitive and unambiguous counting of dividing cells, which is expected to lead to a better reproducibility of the metric.¹⁰⁻¹³ However, it remains little explored in LNETs, particularly in AC.^{9,14}

In this study we take advantage of the multicentre, international lungNENomics series (<https://rarecancersgenomics.com/lungnenomics/>), including 259 LNETs, enriched for AC cases, to assess how whole-slide image (WSI) deep learning analyses, as well as Ki-67 and PHH3 protein

expression could help overcome the current limitations in the histopathological classification of LNETs. Haematoxylin and eosin (HE), Ki-67, and PHH3 stainings for samples of this large cohort were evaluated by six thoracic pathologists, providing an unprecedented opportunity to investigate the reproducibility of the measures. In parallel, WSI deep learning algorithms were developed and applied on the same slides, following both supervised and unsupervised learning paradigms, to extend the pathologists' observations to the whole-slide scale, and to explore potential new morphological features so far unseen by the pathologist's eye which may harbour diagnostic and prognostic value.

PATIENTS AND METHODS

Presentation of the cohort and pathological review

The lungNENomics series is a multicentre, international, retrospective cohort of 259 patients diagnosed with an LNET, whose clinical data are summarised in [Table 1](#). Participants' samples underwent histopathological and deep learning-assisted pathology review for this study ([Figure 1A](#)). A panel of six thoracic pathologists from Italy (MP and GP), France (SL, JMV and AML), and Austria (LB) diagnosed all the cases according to the 2021 WHO guidelines. Mitoses were counted on haematoxylin/eosin (HE)/HE/saffron (HES) sections based on a minimum of three areas of 2 mm². In addition, each reviewer had to analyse subsequently Ki-67 and PHH3 stainings to estimate the expression of these proteins in hotspot areas ([Supplementary Tables S1-S3](#), available at <https://doi.org/10.1016/j.esmoop.2024.103591>). This study was approved by the Ethics Committee of the International Agency for Research on Cancer (IEC Project No. 19-07).

Deep learning-based analyses

Two supervised deep learning algorithms, based on Pathonet models,¹⁵ were independently trained to estimate the proportion of Ki-67 and PHH3-positive cells, per 1000 and per 10 000 detected cells, respectively. To exclude most of the normal areas from the evaluation, tumour areas were previously extracted using CFlow anomaly detection model,¹⁶ as proposed in the study by Mathian et al.¹⁷ Pathonet classified each cell as positive or negative for the marker, allowing marker expression to be estimated at the WSI scale, as opposed to pathologists who assessed protein expression in hotspot areas. The location of the positive cells was then used to calculate spatial metrics to

Table 1. Clinical description based on reference diagnosis

Feature	Pathological review type				Total
	Typical		Atypical		
	<i>n</i>	%	<i>n</i>	%	<i>n</i>
Sex					
Male	72	34.6	28	56.0	100
Female	136	65.4	22	44.0	159
Age, years					
Mean	57		57		57
Median	60		61		60
Range	18-83		22-84		18-84
Location					
Proximal	51	24.5	11	22.0	62
Distal	49	23.6	9	18.0	59
Metastasis	1	0.5	0	0.0	1
Unreported	107	51.4	30	60.0	137
Stage					
IA	89	42.8	14	28.0	103
IB	29	13.9	12	24.0	41
IIA	6	2.9	3	6.0	10
IIB	24	11.5	8	16.0	32
IIIA	13	6.3	6	12.0	19
IIIB	2	1.0	1	2.0	3
IV	3	1.4	0	0.0	3
Unreported	42	20.2	6	12.0	48
Surgery type					
Wedge	12	5.8	1	2.0	13
Segmentectomy	5	2.4	2	4.0	7
Lingulectomy	1	0.5	0	0.0	1
Lobectomy	165	79.3	34	68.0	200
Bilobectomy	5	2.4	1	2.0	6
Pneumonectomy	7	3.4	2	4.0	9
Unreported	13	6.3	10	20.0	23
Post-operative treatment					
None	169	81.3	41	82.0	211
Somatostatin analogue	1	0.5	0	0.0	1
Chemotherapy	0	0.0	2	4.0	2
Chemotherapy and somatostatin analogue	1	0.5	0	0.0	1
Radiotherapy	2	1.0	1	2.0	3
Radiotherapy and chemotherapy	1	0.5	0	0.0	1
Unreported	34	16.3	6	12.0	40
Tobacco smoking history					
Never	81	38.9	20	40.0	102
Former	45	21.6	15	30.0	60
Current	43	20.7	6	12.0	49
Unreported	39	18.8	9	18.0	48
Cannabis use					
Yes	0	0.0	1	2.0	1
No	69	33.2	14	28.0	83
Unreported	139	66.8	35	70.0	175
Other exposures					
None	39	18.8	9	18.0	48
Asbestos	5	2.4	3	6.0	8
Lead, other metals	1	0.5	1	2.0	2
Tar	1	0.5	0	0.0	1
Hair spray, colourants	1	0.5	0	0.0	1
Unspecified	2	1.0	0	0.0	2
Unreported	159	76.4	37	74.0	197
Neuroendocrine genetic disorder					
Yes	0	0.0	3	6.0	3
No	115	55.3	17	34.0	133
Unreported	93	44.7	30	60.0	123
History of cancer					
Yes	37	17.8	4	8.0	41
No	145	69.7	42	84.0	188
Unreported	26	12.5	4	8.0	30
History of radiotherapy					
Yes	8	3.8	0	0.0	8
No	169	81.3	35	70.0	205
Unreported	31	14.9	15	30.0	46
					Continued

*Continued***Table 1. Continued**

Feature	Pathological review type				Total
	Typical		Atypical		
	<i>n</i>	%	<i>n</i>	%	<i>n</i>
Tumour recurrence					
Yes	5	3.8	7	20.0	18
No	120	58.2	27	60.0	152
Unreported	83	38.0	16	20.0	89
Census status					
Alive	202	97.1	43	86.0	245
Dead	4	2.4	7	14.0	12
Unreported	1	0.0	0	0.0	1

see if the spatial patterns varied depending on the type of LNET, as proposed by Bulloni et al.¹⁸

We decided to apply an unsupervised learning algorithm, centred on the Barlow Twins¹⁹ and adapted to WSIs by Quiros et al.,²⁰ to try to discover new morphological features associated with typical or atypical tumour types, and thus not to limit our analysis to criteria already included in the classification. Barlow Twins¹⁹ was trained on a subset of fragments of WSIs of 100 µm² with the aim of generating similar vectors for similar tiles and vice versa for dissimilar tiles, according to the contrastive learning paradigm. The inferred low-dimensional representations of the 4.1 million tiles were then clustered using Leiden community search^{19,20} to identify groups of tiles sharing similar morphological features. Finally, random forest models were used to predict patient diagnosis based on the proportion of tiles belonging to each community within a WSI.²⁰

Statistical framework

Cohen's Kappa coefficient²¹ and Pearson's correlations were calculated to assess agreement between pairs of pathologists and between pathologists and deep learning measurements. This enabled us to identify readers 2 and 6 as outliers in PHH3 measurements compared to others, leading to their exclusion from the corresponding analyses (Supplementary Figure S1A and B, available at <https://doi.org/10.1016/j.esmoop.2024.103591>).

Hypothetical classification systems were created by replacing the mitotic count criteria with the measured expression of Ki-67 and PHH3, with nine threshold values chosen for each marker to explore the effect of thresholds, which is often debated.^{8,14} The classification's relevance was assessed based on the reproducibility of the diagnoses and the prognostic value of the two groups obtained, by majority voting, according to recurrence-free survival (RFS) after surgery, which considers patients who have relapsed or died of the disease or from an unreported cause. To assess whether the new classification systems defining 'new groups' had similar or different prognostic values to the official classification, we measured the differences in prognosis between the new groups (Cox model 1); we also compared prognosis between the new ACs and consensus ACs (Cox model 2). To determine whether the new systems are more sensitive in

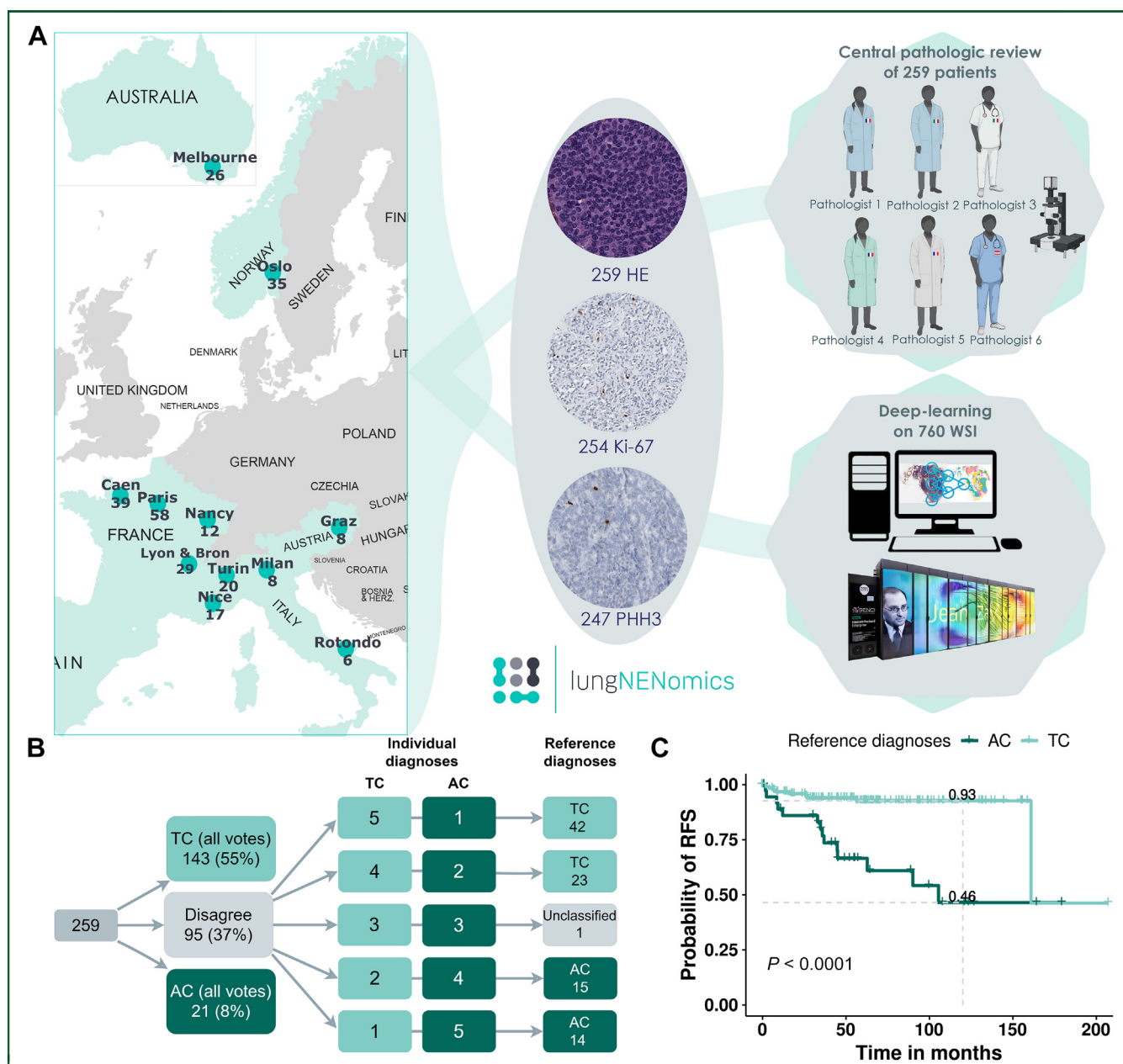


Figure 1. Presentation of the lungNENomics cohort and protocols for pathological review and digital pathology analyses. (A) Study design (created with Bio-Render.com). (B) Flow diagram of the central pathological review leading to final diagnoses. (C) Kaplan–Meier curves of RFS according to the reference diagnoses. Grey dashed lines indicate median survival; P value corresponds to the log-rank test.

AC, atypical carcinoid; HE, haematoxylin and eosin; PHH3, phospho-histone H3; RFS, recurrence-free survival; TC, typical carcinoid; WSI, whole-slide image.

detecting cases with a higher risk of relapse, we compared the RFS data between TCs reclassified as ACs according to the new rules and those remaining classified as TCs (Cox model 3).

To assess these hypothetical scenarios at the level of individual pathologists, which is more representative of common clinical practice, the 10-year RFS rate was reported for each diagnosis type according to the majority vote and at the reader level. Mixed Cox models incorporating pathologists' individual observations were calculated to estimate the marginal effect of the hypothetical classifications, and to compare multivariable models. More details are provided in [Supplementary Methods](https://doi.org/10.1016/j.esmoop.2024.103591), available at <https://doi.org/10.1016/j.esmoop.2024.103591>.

RESULTS

Limitations of the current morphological criteria

Out of the 259 cases initially diagnosed as LNET, 143 were unanimously classified as TC and 21 as AC by the six pathologists (Figure 1B). This represents 68.8% of the 208 cases with a reference diagnosis (majority vote based on WHO criteria) of TC, and 42% of the 50 with a reference diagnosis of AC. AC are therefore more likely to be misdiagnosed than TC (Fisher's exact test $P = 5e-4$)² (Figure 1B). One case was not classified due to disagreement over mitoses counting criteria. The overall Kappa Fleiss score was 0.56; however, Kappa scores were highly

variable depending on the pair of readers compared, ranging from 0.38 to 0.78 (Supplementary Figure S2A, available at <https://doi.org/10.1016/j.esmoop.2024.103591>). Individual pathologists' diagnoses were considered to be incorrect if they differed from the reference one. To explain these misclassifications, we examined the interobserver variability between the two criteria used for classification: the number of mitoses and the presence/absence of necrosis. Focal necrosis was observed in 14%-36% of cases with a reference diagnosis of AC, corresponding to 6-15 patients depending on the reader. Thus, the overall Kappa score of 0.52 for observation of necrosis represents moderate agreement for this feature (Supplementary Figure S2B, available at <https://doi.org/10.1016/j.esmoop.2024.103591>). For the number of mitoses on HE/HES, the mean Pearson's correlation score between readers was 0.61 and ranged from 0.48 to 0.89 (Supplementary Figure S2C, available at <https://doi.org/10.1016/j.esmoop.2024.103591>). The distribution of this variable differed significantly between readers (analysis of variance [ANOVA] $P < 2e-16$) (Figure 2A and B). As necrotic foci were rarely observed, most misclassifications would then be explained by the low reproducibility of mitotic counts on HE/HES.

The value of a classification system lies in its ability to specifically identify patients with a poorer prognosis, as this requires tailored clinical management. However, it has previously been shown that the current classification system for LNETs is imperfect at identifying patients at high risk of progressing towards a more aggressive disease.^{22,23} In the lungNENomics series, RFS data were available for 171 patients, with a median follow-up of 56 months. For patients with a reference diagnosis of TC, the RFS rates at 2, 5, and 10 years are 95.8%, 92.7%, and 92.7%, respectively, compared to 86.0%, 66.6%, and 46.5% for ACs, respectively (Figure 1C, Supplementary Table S4, available at <https://doi.org/10.1016/j.esmoop.2024.103591>, for detailed data on cases with event). Nine TCs (4.3%) and 14 ACs (38.9%) relapsed or died of the disease, in the first 10 years following surgery (log-rank test $P = 2e-9$), with these numbers varying when considering individual pathologist's diagnoses. Given the importance of mitotic count to decide on the diagnosis, we investigated whether shifting the mitotic count threshold would create TC and AC groups that better fit the prognosis of the patients. For all the different thresholds tested (from one to nine mitoses), AC always showed worse prognosis than TC but none of the thresholds generated a group of AC with worse prognosis than the reference group (Figure 2C). It is noteworthy that LNETs are not only rare cancers, but also have a relatively good prognosis, thus the number of patients with an event in our large cohort (23/171) is relatively small, limiting the statistical power (Supplementary Figure S3, available at <https://doi.org/10.1016/j.esmoop.2024.103591>).²⁴ Nevertheless, these data also suggest that increasing the size of the cohort may not solve this cut-off problem, as a wide range of cut-offs allows for both specific and sensitive detection of cases associated with an event (Supplementary Figure S4, available at <https://doi.org/10.1016/j.esmoop.2024.103591>). This finding suggests that

there may be limited utility in debating changes to mitotic count thresholds.

Altogether, the results in the large lungNENomics series confirm the limitations of the current classification criteria that have already been pointed out over the past years in single-centre or smaller series.^{2,25,26} Our data also show that it is not through multiple pathologist reviews or changing the threshold for the number of mitoses to define ACs that will substantially improve the prognostic value of such classification. The question now is whether emerging criteria such as Ki-67 and PHH3 expression could bring something new to the table.

Added value of Ki-67 expression assessment in the evaluation of proliferative activity

The expression levels of Ki-67 were assessed by the six pathologists on 253 samples of the lungNENomics series. We found a strong correlation between average mitotic counts on HE/HES and percentage of Ki-67-positive cells, indicating the high performance of pathologists' measurements ($r = 0.68$, $P < 2e-16$) (Supplementary Figure S5A, available at <https://doi.org/10.1016/j.esmoop.2024.103591>). As expected, ACs showed higher proliferative activity than TCs, with median Ki-67 indices of 10% and 2.5%, respectively (Supplementary Figure S5B, available at <https://doi.org/10.1016/j.esmoop.2024.103591>). Despite the high correlation coefficient between pathologists, with a mean coefficient of 0.79 (Supplementary Figure S5C, available at <https://doi.org/10.1016/j.esmoop.2024.103591>), the distribution of Ki-67 indices differed significantly between readers (ANOVA $P = 7e-5$), as illustrated by the median values that varied between 2% and 4% depending on the pathologist (Supplementary Figure S5D, available at <https://doi.org/10.1016/j.esmoop.2024.103591>).

Based on a Ki-67 index cut-off of 20%, two ACs show the morphological features expected for the LNET grade 3 emerging subtype⁶ (Supplementary Figure S6, available at <https://doi.org/10.1016/j.esmoop.2024.103591>). These two ACs show one and seven mitotic counts, respectively, based on the average counts of the six readers. Five TCs have a percentage of Ki-67-positive cells $>10\%$; two out of the three with RFS data available relapsed within the first year after diagnosis, further supporting the prognostic value of Ki-67 expression levels²⁷⁻²⁹ (Supplementary Figure S7A, available at <https://doi.org/10.1016/j.esmoop.2024.103591>). However, according to this hypothetical classification system, the AC group remains unspecific for cases with poor RFS (Supplementary Figure S7B and C, available at <https://doi.org/10.1016/j.esmoop.2024.103591>). Similarly to what we observed for the mitotic count, there was no clear cut-off that best separates LNETs based on prognosis, but rather a range of values with different sensitivity/specificity ratios. Any threshold between 5% and 17% of Ki-67-positive cells allowed us to define two groups that had different prognosis, but with no difference in prognosis between the Ki-67-based ACs and the reference AC group (Figure 3A and E). Thresholds of 5% and 6% identified a small group of 16 and

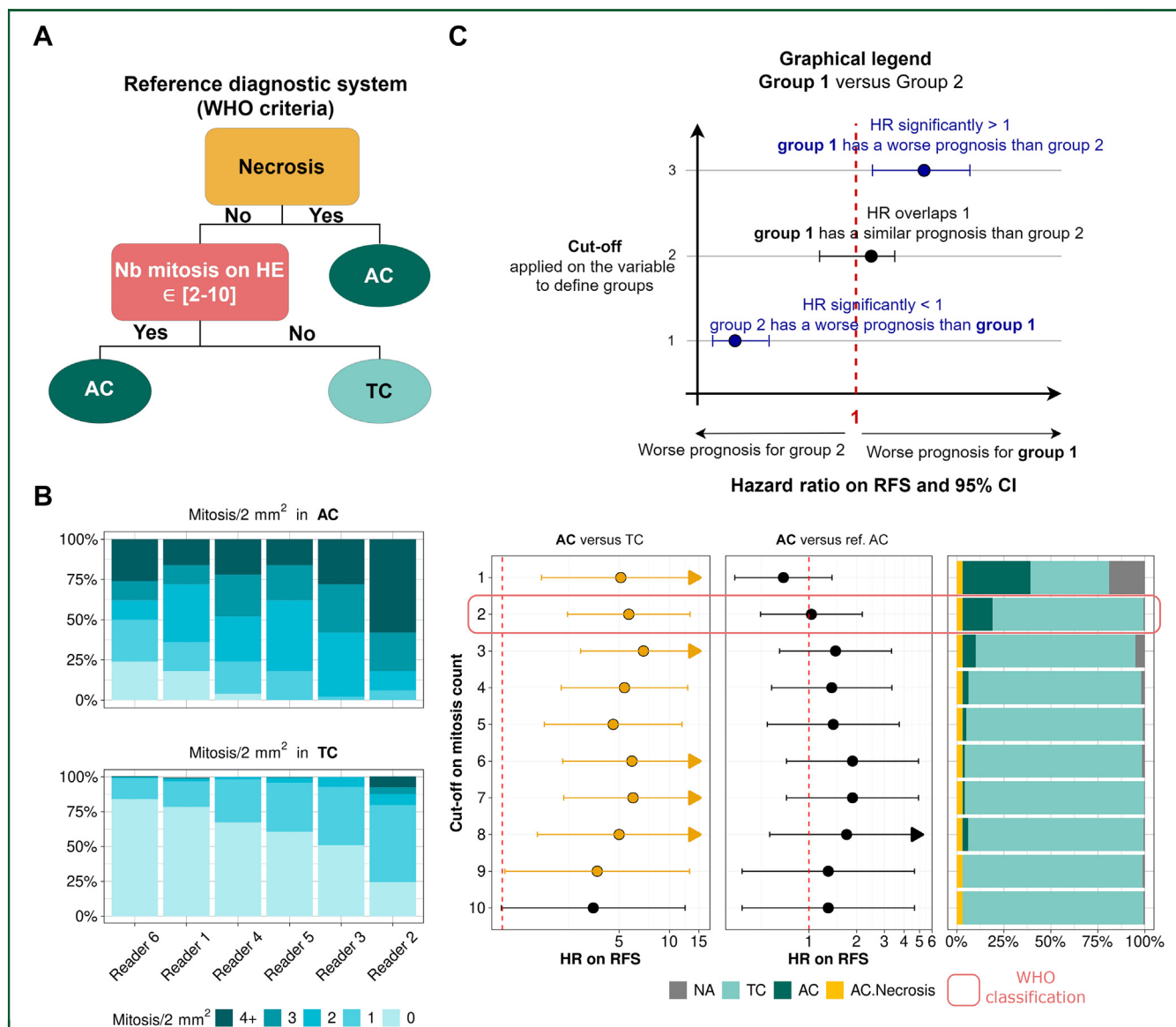


Figure 2. Assessment of number of mitoses counted on HE/HES slides for the classification of LNETs. (A) Reference WHO LNET classification system. (B) Distributions of the number of mitoses per 2 mm² by reader based on reference diagnosis. (C) Prognostic value of TC and AC groups for different mitotic count thresholds. Left panel: forest plot for AC versus TC groups based on RFS. Elements are coloured if the Wald test is significant ($P < 0.05$). The error bars around the hazard ratios correspond to the 95% CI. When the error bars end in an arrow, this means that the confidence intervals are associated with a large value. For ease of reading, this value is not shown and the arrow indicates that the true value is to the right. The name of the group in bold in the panel titles is the target group for which hazard ratios are reported in comparison to the reference group, which is not written in bold. Middle panel: forest plot for AC versus reference AC based on PFS. Right panel: percentages of cases diagnosed as TC and AC. AC cases diagnosed as such because of the presence of necrotic foci are labelled as AC.Necrosis. The proportion of NA represents cases with no majority vote for TC or AC (three votes each).

AC, atypical carcinoid; CI, confidence interval; HE, haematoxylin and eosin; HES, haematoxylin, eosin and saffron; LNET, lung neuroendocrine tumours; NA, not available; Nb, number; PFS, progression-free survival; RFS, recurrence-free survival; TC, typical carcinoid; WHO, World Health Organization.

6 reference TCs, now reclassified as ACs, respectively, with worse prognosis than TCs based on the same thresholds. Similar results were seen in the multivariate analysis (Supplementary Table S5, available at <https://doi.org/10.1016/j.esmoop.2024.103591>). Of note, despite a cut-off of 5% Ki-67-positive cells improving the mean Kappa score between pathologists by 2 points, this does not translate into a significant improvement of the reproducibility (Supplementary Figure S8, available at <https://doi.org/10.1016/j.esmoop.2024.103591>). For all readers, this cut-off identifies the group of TCs at higher risk of relapse, while

the prognostic value of the AC group remains identical to the reference group (Figure 3E).

When applying a deep learning algorithm to Ki-67 expression assessment in WSI, the good correlation between manually and automatically measured Ki-67 indices proved the reliability of the algorithm ($r = 0.77$, $P < 2e-16$), although a different density between AC and TC cannot be proved (Supplementary Figure S9A and B, available at <https://doi.org/10.1016/j.esmoop.2024.103591>). This automatic variable confirmed the higher proliferative activity of reference ACs, with a median index of 16‰ versus 6‰ for TCs. Despite the

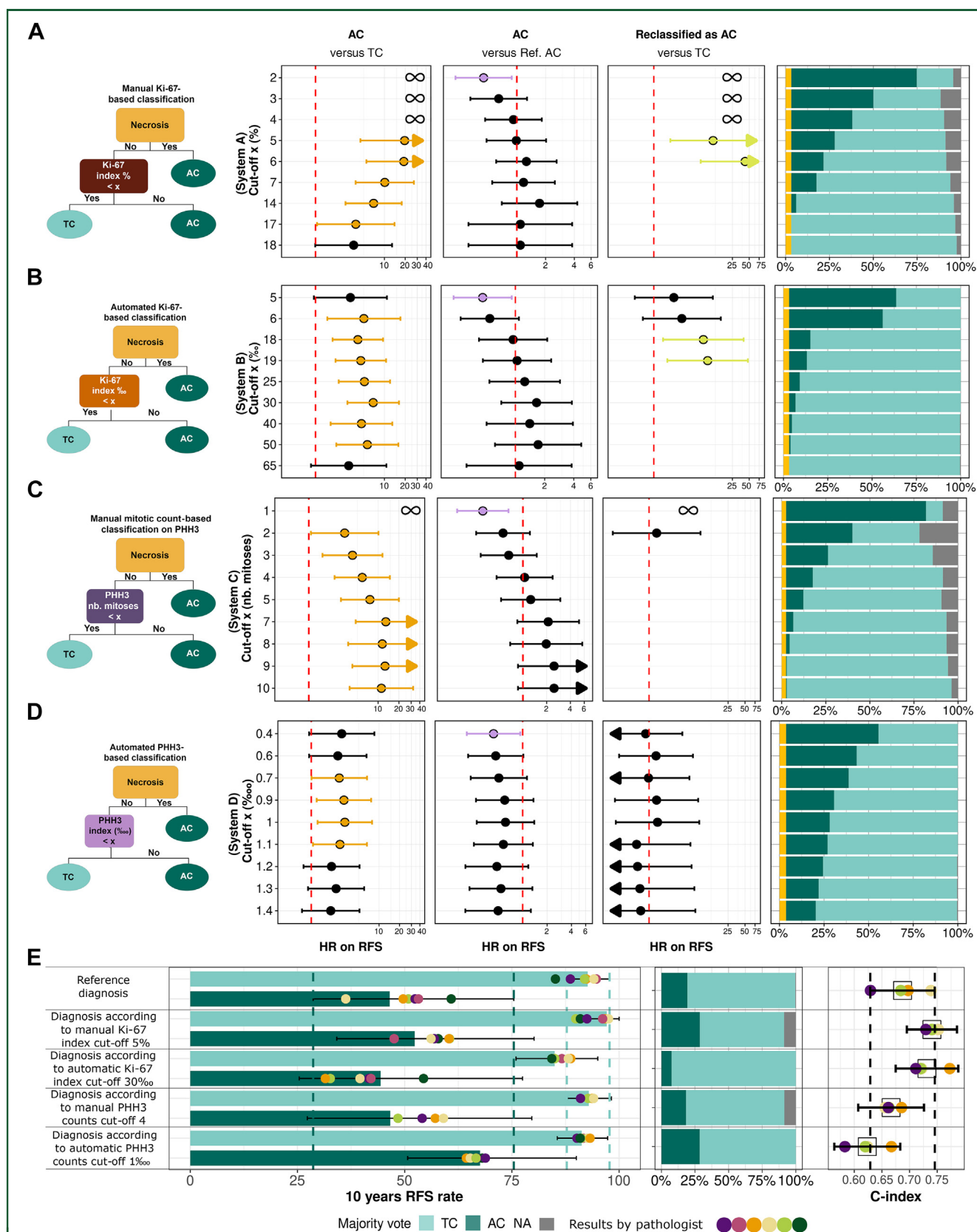


Figure 3. Effect of including Ki-67 or PHH3 assessment in the LNET classification system. First column: hypothetical classification system. Individual pathologists' diagnoses were combined by majority vote, meaning that some cases may remain unclassified if observations resulted in three AC and three TC diagnoses. The four hypothetical classification systems are the following: (A) Based on manual assessment of Ki-67 expression. (B) Based on automatic assessment of Ki-67 expression. (C) Based on manual mitotic count on regions with PHH3 expression. (D) Based on automatic assessment of PHH3 expression. Second to fourth columns: effects of the hypothetical classification system on the prognostic value of the two resulting groups. The significance of the Wald tests associated with the Cox models described on the left is colour-coded for the different thresholds of the variable studied. Fifth column: percentages of cases diagnosed as TC or AC according to the hypothetical classification system. AC cases with foci of necrosis diagnosed by more than three pathologists were labelled AC.Necrosis. The name of the group in bold in the panel

different scaling due to the way Ki-67 is automatically measured on the WSI, instead of counting in hotspots as is done manually by pathologists, the results observed for different thresholds led to similar conclusions (Figure 3B).

Overall, these results suggest that while specific thresholds of Ki-67 expression might be useful for the identification of the few TC cases with higher risk of relapse, the AC cases at a given threshold remain highly unspecific of aggressiveness (Figure 3A, B and E). Therefore, the added value of Ki-67 expression assessment to the current WHO classification is limited. In terms of automated measurement of this marker, time-effective deep learning algorithms could be confidently used by pathologists to quantify Ki-67; however, it would not add any additional information to what pathologists already measure within hotspots, meaning that the region assessed would have minimal impact on the results. As Ki-67 is a broader marker of cell proliferative activity in which mitotic stage is included, it seems pertinent to analyse now whether a more precise a priori detection of mitoses via PHH3 would lead to better results.

Added value of PHH3 expression for the counting of mitotic figures

Similar to Ki-67, PHH3 expression levels were also assessed by immunohistochemistry (IHC) on 246 of the samples of the lungNENomics series. Despite differences in distribution between readers (Supplementary Figure S1A and B, available at <https://doi.org/10.1016/j.esmoop.2024.103591>), we also found a strong correlation between mean mitotic count on HE/HES and on areas with PHH3 expression (PHH3+), demonstrating the quality of the marker ($r = 0.83$, $P < 2e-16$) (Supplementary Figure S10A, available at <https://doi.org/10.1016/j.esmoop.2024.103591>). As already reported,¹⁴ the number of mitoses counted on PHH3+ areas was higher (1.8 times) than on HE/HES. As expected, reference ACs have higher count of mitotic figures on PHH3+ areas than TCs, with a median value of four mitoses versus one for TCs (Supplementary Figure S10B, available at <https://doi.org/10.1016/j.esmoop.2024.103591>). The two putative LNET grade 3 based on Ki-67 showed 10 and 2 mitoses based on PHH3, respectively, further confirming that Ki-67 and PHH3 (and consequently mitotic counts) are not measuring the same feature (Supplementary Figure S6, available at <https://doi.org/10.1016/j.esmoop.2024.103591>). Similar to the results observed for Ki-67, thresholds between 2 and 10 mitoses counted on PHH3+ areas defined TC and AC groups with different prognosis but with the prognosis of the AC group being similar to that of the reference AC (Figure 3C and E). We note that although the

threshold of four mitoses implied little change in classification, it improves the Kappa index by 1 point (Supplementary Figure S8, available at <https://doi.org/10.1016/j.esmoop.2024.103591>). The multivariate Cox model including PHH3 fit the RFS data better than a model based on reference criteria alone (Supplementary Table S5, available at <https://doi.org/10.1016/j.esmoop.2024.103591>).

Regarding the automatic counting of PHH3, there was as expected a correlation between the mean number of mitoses on PHH3+ areas counted manually and automatically, but a drop in performance compared with Ki-67, probably linked to the scarcity of positive cells ($r = 0.64$, $P < 2e-16$) (Supplementary Figure S10C, available at <https://doi.org/10.1016/j.esmoop.2024.103591>). In PHH3 WSIs, as in Ki-67 WSIs, the pattern of proliferative activity does not provide sufficient evidence to support the idea that ACs are more prone to areas of high mitotic density than TCs (Supplementary Figure S10D, available at <https://doi.org/10.1016/j.esmoop.2024.103591>). Surprisingly, counting mitoses on PHH3+ areas automatically failed to identify the TC group with a worse prognosis suggesting that manual counting would be better in the case of PHH3 (Figure 3D and E).

The above-mentioned results suggest that once the techniques are established in the laboratory and the pathologists are familiar with the deep learning automatic counting algorithms, the automatic assessment of Ki-67 and PHH3 could reduce the time that a given pathologist would spend in making a diagnosis, but would not help to significantly improve the clinical meaningfulness of the current WHO classification criteria. The remaining question is whether deep learning could uncover novel clinically meaningful morphological features beyond those current and emerging that might have escaped the pathologist's eye.

Using deep learning to uncover novel clinically relevant morphological features

Applying a state-of-the-art unsupervised deep learning algorithm²⁰ to 257 HE/HES WSIs failed to separate the two histological types of LNETs, as evidenced by the distribution of tiles (subparts of WSIs) in the two-dimensional map of morphological features (Figure 4A). To enhance model interpretation and bolster its predictive capabilities, we grouped tiles into communities, which are expected to consist of tiles with similar morphological features (Figure 4B; the full description of communities and their associated features is provided in Supplementary Table S6, available at <https://doi.org/10.1016/j.esmoop.2024.103591>).

titles is the target group for which hazard ratios are reported in comparison to the reference group, which is not written in bold. (E) Left panel: For each classification system, the coloured bar corresponds to the 10-year RFS rate based on the diagnoses resulting from majority voting; these rates are associated with the 95% confidence intervals. Middle panel: percentages of cases diagnosed as TC or AC according to the hypothetical classification system resulting from majority voting. Right panel: Harrell's C-index, also known as the correspondence index, comparing the quality of univariate Cox models for RFS incorporating the different diagnoses resulting from the classification system mentioned on the left. Error bars around the estimator correspond to confidence intervals. Vertical lines indicate the boundary of the confidence interval resulting from the reference classification system used for comparison. Coloured dots indicate the results obtained from Cox models built on the diagnoses resulting from each pathologist's observations (one model for each pathologist).

AC, atypical carcinoid; HR, hazard ratio; LNET, lung neuroendocrine tumours; PFS, progression-free survival; PHH3, phospho-histone H3; RFS, recurrence-free survival; TC, typical carcinoid; NA, not available.

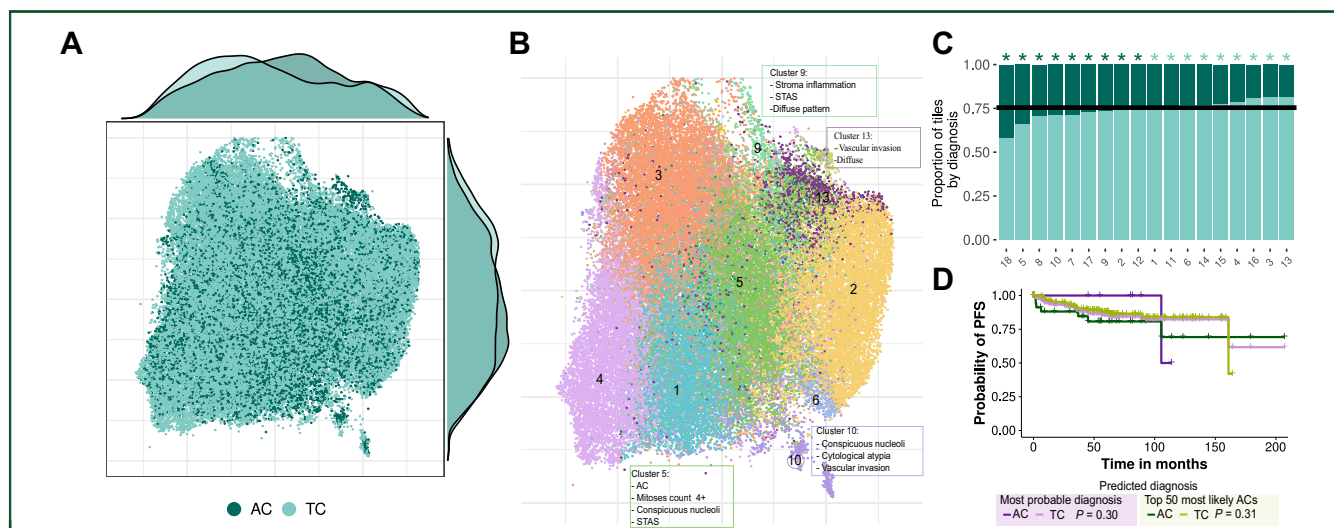


Figure 4. Unsupervised deep learning experiment on HE/HES WSI of LNET patients. (A) Two-dimensional morphological map resulting from the uniform manifold approximation and projection (UMAP) dimensionality reduction technique⁴⁵ applied to Barlow Twins-encoded vectors, each tile is coloured according to the reference diagnosis of the patients. (B) Representation of the 18 communities on (A). Some communities are annotated according to enrichment for certain morphological features, in line with the annotations of pathologists on the WSI. (C) Proportion of tiles in each Leiden community by tumour type. The horizontal black line represents the total proportion of TC versus AC tiles included in the experiment. At the top of the bar, the presence of a star indicates whether a community is significantly enriched for a type; the colour of the star indicates for which tumour types it is enriched. (D) Kaplan–Meier curves of PFS as a function of predicted diagnoses using random forest. The purple curves correspond to the diagnoses with the highest probability between the two types. The green curves correspond to the diagnoses predicted if the 50 most likely ACs were classified as such to obtain a group of the same size as the reference diagnosis. The *P* values of the log-rank test associated with the type of predictions are shown in the legend.

AC, atypical carcinoid; HE, haematoxylin and eosin; HES, haematoxylin, eosin and saffron; LNET, lung neuroendocrine tumours; STAS, spread through alveolar spaces; TC, typical carcinoid; WSI, whole-slide image.

591). The random forest trained on the proportions of tiles in each of the 18 communities calculated per patient allowed us to predict the reference diagnosis with very low accuracy (Figure 4C). Indeed, if we took the highest probability between the two classes as the predicted diagnosis, we obtain a receiver operating characteristic (ROC) score of only 0.59; a score equal to or greater than this was obtained by chance 19 times out of 500 permutation tests (permutation test $P = 0.04$) (Supplementary Figure S11A and B, available at <https://doi.org/10.1016/j.esmoop.2024.103591>). The Kappa score between predicted and reference diagnoses is much lower than that obtained by pathologists, with a value of 0.15, in comparison to 0.56. Interestingly, several communities were enriched for specific morphological features, described at the WSI level, for example spread through alveolar spaces (STAS) and vascular invasion (Figure 4B; Supplementary Table S6, available at <https://doi.org/10.1016/j.esmoop.2024.103591>). However, the distribution of morphological features described at the WSI level within communities does not show a clear association with the reference diagnosis (Supplementary Table S6, available at <https://doi.org/10.1016/j.esmoop.2024.103591>), as highlighted by the statistics on additional features (STAS, vascular invasion, etc.) reported by pathologists (Supplementary Figure S12, available at <https://doi.org/10.1016/j.esmoop.2024.103591>). In fact, none of these characteristics were significantly associated with one of the two histological types, according to the summary of the majority vote (Supplementary Figure S12, available at <https://doi.org/10.1016/j.esmoop.2024.103591>). The validity of our implementation of the models of Quiros and colleagues²⁰

was demonstrated by a respectable performance in the classification of lung adenocarcinomas from lung squamous cell carcinoma in a set of WSIs available from The Cancer Genome Atlas (TCGA) Program database, with a ROC score of 0.93 (Supplementary Figure S13A–C, available at <https://doi.org/10.1016/j.esmoop.2024.103591>). As shown in Figure 4D, the TC and AC groups predicted by deep learning have a similar prognosis, whether the groups are defined according to the highest probability, or ACs are defined as the 50 most probable cases (to obtain a group of the same size as the reference diagnosis).

DISCUSSION

The current WHO classification of LNETs established according to the number of mitotic figures counted on HE slides, and the presence of necrotic foci, is moderately reproducible as previously suggested³⁰ and as confirmed in our large, multi-centric, and international lungNENomics series. This interobserver variability is mainly explained by the low reproducibility on the counting of mitotic figures, given that the presence of necrosis represents a marginal feature only present in a minority of ACs. Our series also confirms the limitations of the current classification in predicting prognosis as shown by the up to 11% of TC patients who had an event within the first year after resection of the primary tumour (13/121), and the up to 73% of ACs who did not have an event after 10 years (45/62) (Figure 3E).

The interobserver variability associated with mitotic counting on HE has a strong impact on diagnosis. In addition, the limitations of the current morphological criteria to

specifically identify aggressive cases translates in longer follow-up for all patients with an LNET, which is costly and difficult for the patient, creating a lot of anxiety. The systematic reporting of Ki-67 index has been encouraged by the European and North American NET societies^{31,32} but the debate over the inclusion of the Ki-67 index in the classification of LNETs (as used for gastroenteropancreatic NET since 2010) has revolved around the optimal threshold for distinguishing TC and AC.^{4,8,28,33-35} Here we demonstrate that there is a wide range of cut-offs of Ki-67 expression that allow LNETs to be divided into two groups with similar prognostic values to those defined by the current WHO criteria. And while the use of Ki-67 seems to improve the identification of some TCs with higher risk of relapse, demonstrating that intense proliferative activity should redirect the diagnosis towards the AC type,^{6,36,37} the use of this marker does not improve the specificity of the group of ACs to aggressive cases. In addition, we have shown that Ki-67 does not reduce interobserver variability, despite the fact that staining was carried out centrally, which has previously been shown to reduce inter-laboratory variability in Ki-67.³⁸ Similarly, mitoses can be counted on PHH3+ areas without a major improvement when compared to the current criteria, for a wide range of thresholds. This is particularly true for a threshold of four mitoses, which is consistent with the results of the studies by Tsuta et al., Kim et al., and Laflamme et al.^{9,13,14} While PHH3 reduces the time required by the pathologist to count mitoses,^{10,11,14} we cannot prove that it significantly reduces interobserver variability, as previously claimed.^{10,14} Finally, our data on the automatic evaluation of Ki-67 and PHH3 by supervised deep learning algorithms show that these algorithms reach the performance of expert pathologists in terms of prognostic value, as they allow division of LNET into two groups with a similar prognostic value to that defined by the experts, they mechanically reduce the interobserver variability, and are widely recognised as being time efficient. However, they do not provide any additional information to the manual assessment, either for Ki-67 or PHH3. Overall, if these markers were to be adopted, optimal thresholds should be defined on the basis of the best clinical strategy that can be implemented for patients, taking into account the costs, benefits, and risks of over- or under-diagnosis of AC, and not simply on the basis of RFS or overall survival data, as these prognostic data are inherent to the cohort and may not represent the 'true' distributions, given the low frequency of events among such rare cancers. Given the degree of interobserver variability in detecting foci of necrosis, it would be interesting to investigate the potential added value of expression of HIF-1-alpha, a protein that is expressed under hypoxic conditions and localised close to the necrotic area.³⁹

Finally, applying unsupervised state-of-the-art deep learning methods to WSIs does not help in distinguishing TC from AC, given the great histological similarity between these entities. This suggests that we might have reached a plateau on what morphology can bring to the distinction

between TCs and ACs and, more importantly, to predicting aggressive disease in LNET patients. On the other hand, although with some limitations, the prognostic value of the current morphological classification is unarguable, it is not useful for personalised treatments.

Conclusions

While not having an added value in terms of prognostication, the assessment of Ki-67 and PHH3 by IHC could be suggested in the upcoming WHO Classification of Thoracic Tumours to guide the assessment of tumour proliferation. However, our data suggest that efforts should be put elsewhere if we want to make a breakthrough in improving the diagnosis, prognostication, and clinical management of these diseases. There are emerging molecular markers that may complement the current morphological criteria to predict aggressive disease. For example, the prognostic value of CD44, ASCL1, and OTP expression,^{40,41} or *TERT*⁴² has been suggested. Beyond single-molecular markers, we and others have shown through multi-omics data analyses the existence of robust molecular groups that only partially match the separation into TC and AC.^{43,44} These kinds of studies are needed to better understand the biology and aetiology of LNETs and therefore to open new avenues for the clinical management of these rare and understudied diseases, through a more clinically relevant morpho-molecular classification.

ACKNOWLEDGEMENTS

The lungNENomics project is part of the Rare Cancers Genomics initiative (www.rarecancersgenomics.com/) led by the Rare Cancers Genomics team at the IARC (<https://www.iarc.who.int/teams-gem-rcg/>). This work is also part of the European Neuroendocrine Tumor Society (ENETS) lung task force. We thank the Hospices Civils de Lyon (CRB-HCL BB-0033-00046) and Centre Léon Bérard (CRB-CLB BB-0033-00050) biobanks in Lyon, France, both authorised by the French Ministry of Research, for sharing human biological samples and associated data. We thank the 12 centres that voluntarily participated in the lungNENomics project by providing one FFPE block per patient: The Tumour Bank of the François Baclesse Centre in Caen (France), the Institut für Diagnostik und Forschung in Pathologie of the Medical University of Graz (Austria), the Department of Biopathology of the Léon Bérard Centre in Lyon (France), the Institute of Pathology of the Hospices civils de Lyon (France), the Department of Surgical Oncology of St Vincent's Hospital in Melbourne (Australia), the Department of Oncology and Haemato-Oncology of the University of Milan (Italy), the Department of Biopathology at the Nancy Regional Hospital (France), the Laboratory of Clinical and Experimental Pathology at the Pasteur Hospital in Nice (France), the Departments of Pathology and Oncology at Oslo University Hospital (Norway), the Pathology Department at the Cochin Hospital in Paris (France), the Oncology Unit at the IRCCS Cas Sollievo della Sofferenza Foundation

in Rotondo (Italy), and the Oncology Department at the University of Turin (Italy). We thank the 2 anonymous reviewers for their insightful comments and suggestions, which significantly contributed to the improvement of our manuscript. The results published here are in whole or part based upon data generated by the TCGA Research Network: <https://www.cancer.gov/tcga>. We also acknowledge the contribution of Associate Professor Gavin Wright and Dr Behnoush Abedi-Ardekani.

FUNDING

This work was supported by HPC resources from GENCI-IDRIS [grant numbers 2022-AD011012172R1 and 2024-AD010315173]. This work was also supported by the Neuroendocrine Tumor Research Foundation (NETRF, Investigator Award 2022 to M.F.), Worldwide Cancer Research (WCR) [grant number 21-0005 to L.F.C.], the French National Cancer Institute (INCa) [grant number PRT-K-17-047 to L.F.C.], and LYRICAN+ [grant number INCa-DGOS-INSERM-ITMO cancer_18003].

DISCLOSURE

Where authors are identified as personnel of the International Agency for Research on Cancer/WHO, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy, or views of the International Agency for Research on Cancer/WHO. The rest of the authors declare no conflict of interest.

DATA SHARING

The data used in the current study are available in the ESMOOpen_LungNENomicsCohort repository https://github.com/IARCBioinfo/ESMOOpen_LungNENomicsCohort. The scripts needed to reproduce the deep learning experiments are available at:

- WSIs pre-processing: <https://github.com/IARCBioinfo/WSIPreprocessing>
- Tumour segmentation by anomaly detection: <https://github.com/IARCBioinfo/TumorSegmentationCFlowAD> (based on: <https://github.com/gudovskiy/cflow-ad>)
- Pathonet adapted to LNET: <https://github.com/IARCBioinfo/PathonetLNET> (based on: <https://github.com/SHIDCenter/PathoNet>)
- Barlow Twins for LNET: <https://github.com/IARCBioinfo/LNETBarlowTwins> (based on: <https://github.com/facebookresearch/barlowtwins>)

Access to WSI is available on request to sylvie.lantuejoul@lyon.unicancer.fr.

REFERENCES

1. WHO Classification of Tumours. *Thoracic Tumors (ed 5)*. Lyon, France: IARC Press; 2021.
2. Swarts DRA, van Suylen RJ, den Bakker MA, et al. Interobserver variability for the WHO classification of pulmonary carcinoids. *Am J Surg Pathol*. 2014;38(10):1429-1436.
3. Lee CH, Chang HK, Lee HW, Shin DH, Roh MS. The interobserver variability for diagnosing pulmonary carcinoid tumor. *Korean J Pathol*. 2010;44(3):267.
4. Warth A, Fink L, Fisseler-Eckhoff A, et al. Interobserver agreement of proliferation index (Ki-67) outperforms mitotic count in pulmonary carcinoids. *Virchows Arch*. 2013;462(5):507-513.
5. Pelosi G, Travis WD. The Ki-67 antigen in the new 2021 World Health Organization classification of lung neuroendocrine neoplasms. *Pathologica*. 2021;113(5):377-387.
6. Rekhman N. Lung neuroendocrine neoplasms: recent progress and persistent challenges. *Mod Pathol*. 2022;35(suppl 1):36-50.
7. Rindi G, Mete O, Uccella S, et al. Overview of the 2022 WHO classification of neuroendocrine neoplasms. *Endocr Pathol*. 2022;33(1):115-154.
8. Pelosi G, Rindi G, Travis WD, Papotti M. Ki-67 antigen in lung neuroendocrine tumors: unraveling a role in clinical practice. *J Thorac Oncol*. 2014;9(3):273-284.
9. Tsuta K, Liu DC, Kalhor N, Wistuba II, Moran CA. Using the mitosis-specific marker anti-phosphohistone H3 to assess mitosis in pulmonary neuroendocrine carcinomas. *Am J Clin Pathol*. 2011;136(2):252-259.
10. Voss SM, Riley MP, Lokhandwala PM, Wang M, Yang Z. Mitotic count by phosphohistone H3 immunohistochemical staining predicts survival and improves interobserver reproducibility in well-differentiated neuroendocrine tumors of the pancreas. *Am J Surg Pathol*. 2015;39(1):13-24.
11. Villani V, Mahadevan KK, Ligorio M, et al. Phosphorylated histone H3 (PHH3) is a superior proliferation marker for prognosis of pancreatic neuroendocrine tumors. *Ann Surg Oncol*. 2016;23(suppl 5):609-617.
12. Dumars C, Foubert F, Toucheffeu Y, et al. Can PPH3 be helpful to assess the discordant grade in primary and metastatic enteropancreatic neuroendocrine tumors? *Endocrine*. 2016;53(2):395-401.
13. Kim MJ, Kwon MJ, Kang HS, et al. Identification of phosphohistone H3 cutoff values corresponding to original WHO grades but distinguishable in well-differentiated gastrointestinal neuroendocrine tumors. *Biomed Res Int*. 2018;2018:1-10.
14. Laflamme P, Mansoori BK, Sazanava O, et al. Phospho-histone-H3 immunostaining for pulmonary carcinoids: impact on clinical appraisal, interobserver correlation, and diagnostic processing efficiency. *Hum Pathol*. 2020;106:74-81.
15. Negahbani F, Sabzi R, Pakniyat Jahromi B, et al. PathoNet introduced as a deep neural network backend for evaluation of Ki-67 and tumor-infiltrating lymphocytes in breast cancer. *Sci Rep*. 2021;11(1):8489.
16. Gudovskiy D, Ishizaka S, Kozuka K. CFLOW-AD: real-time unsupervised anomaly detection with localization via conditional normalizing flows [Internet]. *arXiv [cs.CV]*. 2021:98-107 [cited 2022 Dec 19]. Available at: https://openaccess.thecvf.com/content/WACV2022/html/Gudovskiy_CFLOW-AD_Real-Time_Unsupervised_Anomaly_Detection_With_Localization_via_Conditional_Normalizing_WACV_2022_paper.html.
17. Mathian E, Liu H, Fernandez-Cuesta L, Samaras D, Foll M, Chen L. HaloAE: a local transformer auto-encoder for anomaly detection and localization based on HaloNet. In: *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications* [Internet]. SCITEPRESS - Science and Technology Publications; 2023. <https://doi.org/10.5220/0011865900003417>.
18. Bulloni M, Sandrini G, Stacchiotti I, et al. Automated analysis of proliferating cells spatial organisation predicts prognosis in lung neuroendocrine neoplasms. *Cancers (Basel)*. 2021;13(19):4875.
19. Zbontar J, Jing L, Misra I, LeCun Y, Deny S. Barlow twins: self-supervised learning via redundancy reduction. In: Meila M, Zhang T, editors. *Proceedings of the 38th International Conference on Machine Learning*. Vol. 139. PMLR; July 18-24, 2021:12310-12320.
20. Quiros AC, Coudray N, Yeaton A, et al. Self-supervised learning in non-small cell lung cancer discovers novel morphological clusters linked to patient outcome and molecular phenotypes [Internet]. *arXiv [cs.CV]*. 2022. <http://arxiv.org/abs/2205.01931>.
21. Cohen J. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychol Bull*. 1968;70(4):213-220.

22. Dermawan JK, Farver CF. The prognostic significance of the 8th edition TNM staging of pulmonary carcinoid tumors. *Am J Surg Pathol*. 2019;43(9):1291-1296.
23. Fernandez-Cuesta L, Sexton-Oates A, Bayat L, Foll M, Lau SCM, Leal T. Spotlight on small-cell lung cancer and other lung neuroendocrine neoplasms. *Am Soc Clin Oncol Educ Book*. 2023;43:e390794.
24. Schoenfeld DA. Sample-size formula for the proportional-hazards regression model. *Biometrics*. 1983;39(2):499-503.
25. Travis WD, Gal AA, Colby TV, Klimstra DS, Falk R, Koss MN. Reproducibility of neuroendocrine lung tumor classification. *Hum Pathol*. 1998;29(3):272-279.
26. Skov BG, Krasnik M, Lantuejoul S, Skov T, Brambilla E. Reclassification of neuroendocrine tumors improves the separation of carcinoids and the prediction of survival. *J Thorac Oncol*. 2008;3(12):1410-1415.
27. Pelosi G, Massa F, Gatti G, et al. Ki-67 evaluation for clinical decision in metastatic lung carcinoids: a proof of concept. *Clin Pathol*. 2019;12:2632010X19829259.
28. Marchiò C, Gatti G, Massa F, et al. Distinctive pathological and clinical features of lung carcinoids with high proliferation index. *Virchows Arch*. 2017;471(6):713-720.
29. Centonze G, Maisonneuve P, Simbolo M, et al. Lung carcinoid tumours: histology and Ki-67, the eternal rivalry. *Histopathology*. 2023;82(2):324-339.
30. Swarts DR. Interobserver variability for the WHO classification of pulmonary carcinoids. *Am J Surg Pathol*. 2014;38:1429-1436.
31. Singh S, Bergsland E, Card C. CommNETs/NANETS guidelines for the diagnosis and management of patients with lung neuroendocrine tumors: an international collaborative endorsement and update of the 2015 ENETS expert consensus guidelines. *J Thorac Oncol*. 2020;15:1577-1598.
32. Caplin ME, Baudin E, Ferolla P, et al. Pulmonary neuroendocrine (carcinoid) tumors: European Neuroendocrine Tumor Society expert consensus and recommendations for best practice for typical and atypical pulmonary carcinoids. *Ann Oncol*. 2015;26(8):1604-1620.
33. Swarts DRA, Rudelius M, Claessen SMH, et al. Limited additive value of the Ki-67 proliferative index on patient survival in World Health Organization-classified pulmonary carcinoids. *Histopathology*. 2017;70(3):412-422.
34. Garg R, Bal A, Das A, Singh N, Singh H. Proliferation marker (Ki67) in sub-categorization of neuroendocrine tumours of the lung. *Turk Patoloji Derg*. 2019;35(1):15-21.
35. Fabbri A, Cossa M, Sonzogni A, et al. Ki-67 labeling index of neuroendocrine tumors of the lung has a high level of correspondence between biopsy samples and surgical specimens when strict counting guidelines are applied. *Virchows Arch*. 2017;470(2):153-164.
36. Dermawan JKT, Farver CF. The role of histologic grading and Ki-67 index in predicting outcomes in pulmonary carcinoid tumors. *Am J Surg Pathol*. 2020;44(2):224-231.
37. Marchevsky AM, Hendifar A, Walts AE. The use of Ki-67 labeling index to grade pulmonary well-differentiated neuroendocrine neoplasms: current best evidence. *Mod Pathol*. 2018;31(10):1523-1531.
38. Focke CM, Bürger H, van Diest PJ, et al. Interlaboratory variability of Ki67 staining in breast cancer. *Eur J Cancer*. 2017;84:219-227.
39. Daskalakis K, Kaltsas G, Öberg K, Tsolakis AV. Lung carcinoids: long-term surgical results and the lack of prognostic value of somatostatin receptors and other novel immunohistochemical markers. *Neuroendocrinology*. 2018;107(4):355-365.
40. Centonze G, Maisonneuve P, Simbolo M, et al. Ascl1 and OTP tumour expressions are associated with disease-free survival in lung atypical carcinoids. *Histopathology*. 2023;82(6):870-884.
41. Swarts DRA, Henfling MER, van Neste L, et al. CD44 and OTP are strong prognostic markers for pulmonary carcinoids. *Clin Cancer Res*. 2013;19(8):2197-2207.
42. Werr L, Bartenhagen C, Rosswog C, Cartolano M, et al. TERT expression defines clinical outcome in pulmonary carcinoids. In: *J Clin Oncol*. 2024; In Press.
43. Alcalá N, Leblay N, Gabriel AAG, et al. Integrative and comparative genomic analyses identify clinically relevant pulmonary carcinoid groups and unveil the supra-carcinoids. *Nat Commun*. 2019;10(1):3407.
44. Laddha SV, Silva D, Robzyk EM. Integrative genomic characterization identifies molecular subtypes of lung carcinoids. *Cancer Res*. 2019;79(17):4339-4347.
45. McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction [Internet]. *arXiv [stat.ML]*. 2018. <http://arxiv.org/abs/1802.03426>.