nature medicine

LETTERS

https://doi.org/10.1038/s41591-019-0583-3

Deep learning-based classification of mesothelioma improves prediction of patient outcome

Pierre Courtiol[1,8], Charles Maussion[1,8], Matahi Moarii[1], Elodie Pronier[1], Samuel Pilcer[1], Meriem Sefta[1], Pierre Manceron[1], Sylvain Toldo[1], Mikhail Zaslavskiy[1], Nolwenn Le Stang[2], Nicolas Girard[3,4], Olivier Elemento[5], Andrew G. Nicholson[6], Jean-Yves Blay[7], Françoise Galateau-Sallé[2,8], Gilles Wainrib[1,8] and Thomas Clozel[1,8*]

Malignant mesothelioma (MM) is an aggressive cancer primarily diagnosed on the basis of histological criteria[1]. The 2015 World Health Organization classification subdivides mesothelioma tumors into three histological types: epithelioid, biphasic and sarcomatoid MM. MM is a highly complex and heterogeneous disease, rendering its diagnosis and histological typing difficult and leading to suboptimal patient care and decisions regarding treatment modalities[2]. Here we have developed a new approach—based on deep convolutional neural networks—called MesoNet to accurately predict the overall survival of mesothelioma patients from whole-slide digitized images, without any pathologist-provided locally annotated regions. We validated MesoNet on both an internal validation cohort from the French MESOBANK and an independent cohort from The Cancer Genome Atlas (TCGA). We also demonstrated that the model was more accurate in predicting patient survival than using current pathology practices. Furthermore, unlike classical black-box deep learning methods, MesoNet identified regions contributing to patient outcome prediction. Strikingly, we found that these regions are mainly located in the stroma and are histological features associated with inflammation, cellular diversity and vacuolization. These findings suggest that deep learning models can identify new features predictive of patient survival and potentially lead to new biomarker discoveries.

MM is a rare, highly lethal cancer of the pleural lining. It is derived from the lining of the serous cavities and is associated with asbestos exposure in 80% of cases[3,4]. It can be latent for decades and, although strict regulations regarding asbestos use have been defined, its incidence rate is still expected to increase over the next few years[5]. Currently, diagnosis is established by pathologists via tissue biopsy, and patients are classified according to the 2015 World Health Organization (WHO) classification[1] into three conventional histologic types: epithelioid MM (EMM), sarcomatoid MM (SMM) and biphasic MM (BMM), with BMM containing a mix of sarcomatoid and epithelioid components. EMM patients have the most favorable prognosis (median overall survival (OS) of 16 months), whereas SMM patients have the worst prognosis (OS of 5 months) and BMM patients have an intermediate prognosis[6]. This histological classification is of prognostic and therapeutic value[7,8] but is insufficient to cover the extreme variability in clinical features and patient outcomes in MM patients. This highlights the important need for developing new methods to identify predictive biomarkers consistently associated with survival[2,9–11].

The advent of deep learning and the availability of thousands of histology slides provides a new opportunity to revisit classical approaches to diagnosis and predicting patient outcomes[12–21]. However, this approach is usually seen as a black-box, where image features contributing to the prediction are hardly intelligible. To address these limitations, we developed MesoNet, a deep learning algorithm specifically customized to analyze large images, such as whole-slide images (WSIs), without any local annotation by pathologists[22].

To build MesoNet, we adapted a recently described algorithm specifically designed to address this scenario[22]. To create prediction models, our algorithm trains deep learning networks from WSIs[23,24] with only global data labels (Extended Fig. 1). First, WSIs from MM patients were preprocessed and divided into small $112 \times 112\,\mu m$ squares (224×224 pixels), called 'tiles'. These tiles were fed into the network architecture, which assigned a 'survival score' to each tile, through an iterative learning process. Finally, the network selected the tiles of each WSI that were the most relevant to the prediction and used this limited number of tiles to predict patient OS (for a detailed description see Extended Methods).

To train and test MesoNet, we used a dataset of 2,981 slides from 2,981 pleural mesothelioma patients from multiple French institutions (MESOPATH/MESOBANK[25]) (for the detailed patient descriptions see Supplementary Table 1) and randomly divided it into a Train set (2,300 slides) and an entirely separate Test set (681 slides). On the Train set, a fivefold cross-validation strategy was applied. The model was then trained on all 2,300 training slides and evaluated on the Test set of 681 slides. Moreover, a validation set of 56 WSIs from 56 MM patients from an independent dataset was obtained from the TCGA[10] to test the robustness of MesoNet (Fig. 1) (Extended Methods).

[1]Owkin Lab, Owkin, Inc., New York, NY, USA. [2]Department of Biopathology, MESOPATH/MESOBANK Cancer Center Léon Bérard, Lyon, France. [3]Université de Lyon, Université Claud Bernard Lyon 1, Lyon, France. [4]Institut du Thorax Curie-Montsouris, Institut Curie, Paris, France. [5]Department of Physiology and Biophysics, Institute for Computational Biomedicine and Caryl and Israel Englander Institute for Precision Medicine, WorldQuant Initiative for Quantitative Prediction, Weill Cornell Medicine, New York, NY, USA. [6]Department of Histopathology, Royal Brompton and Harefield Hospitals NHS Foundation Trust, and National Heart and Lung Institute, Imperial College, London, UK. [7]Department of Medical Oncology, Centre Léon Bérard, Lyon, France. [8]These authors contributed equally: Pierre Courtiol, Charles Maussion, Françoise Galateau-Sallé, Gilles Wainrib, Thomas Clozel. *e-mail: thomas.clozel@owkin.com

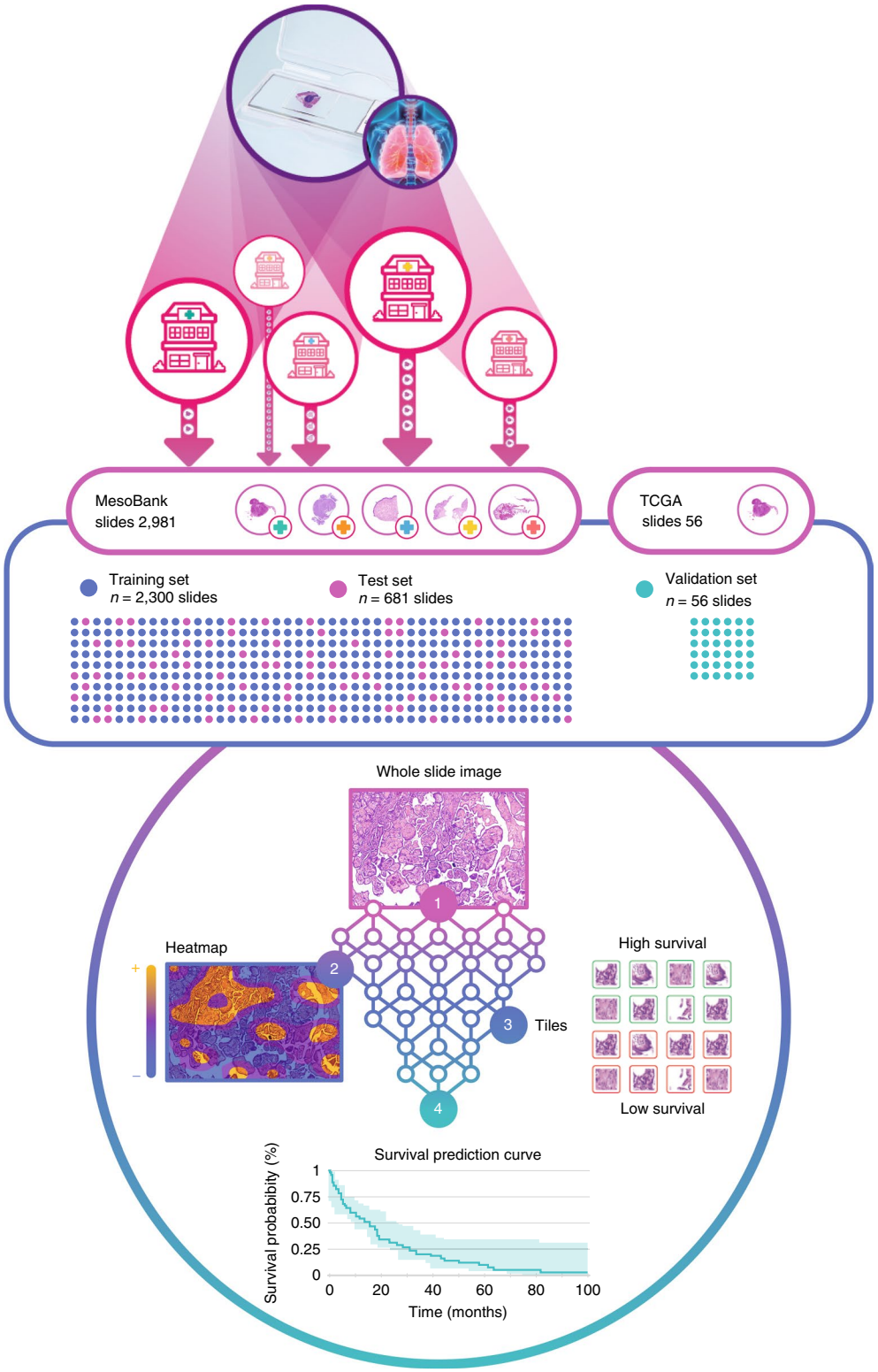


Fig. 1 | MesoNet layout. Mesothelioma histology slides were collected from the MESOPATH/MESOBANK database and The Cancer Genome Atlas (TCGA). First, we trained a machine learning model to predict patient overall survival using 2,300 WSIs randomly chosen from a total of 2,981 slides taken from MESOBANK, without expert-derived annotations. The model was then tested on the remaining 681 slides from MESOBANK and validated on 56 slides from TCGA. A predictive score was given to each tile of interest, positively or negatively associated with survival (see heatmap). Extremal tiles filtering allowed us to retain the more informative tiles for use by the model for prediction so as to achieve patient survival prediction (shaded area over the curve represents CI).

We then compared MesoNet to the pathologist-provided histology type ('Histo'), 'Histo/Grade', a model combining histological type and tumor grade, and 'Meanpool', a naïve machine learning approach (see 'Baseline and other deep learning models' section in the Extended Methods). We used the *c* index to compare the predictive performance of each model (see 'Assessment of the

performance' section in the Extended Methods). MesoNet significantly outperformed the baseline models on the cross-validation set from the Train dataset ($c_{\text{MesoNet}}(\text{Train}) = 0.642$ versus $c_{\text{Histo}}(\text{Train}) = 0.596$, $P < 1 \times 10^{-30}$, t-test), the Test dataset ($c_{\text{MesoNet}}(\text{Test}) = 0.643$ versus $c_{\text{Histo}}(\text{Test}) = 0.598$, $P < 1 \times 10^{-30}$, t-test) and the TCGA dataset ($c_{\text{MesoNet}}(\text{TCGA}) = 0.656$ versus $c_{\text{Histo}}(\text{TCGA}) = 0.590$, $P < 1 \times 10^{-30}$, t-test), demonstrating the effectiveness of the algorithm in predicting patient outcome (Fig. 2a). MesoNet performed slightly less well than Meanpool on the MESOBANK training and test sets ($c_{\text{MesoNet}}(\text{Train}) = 0.642$ versus $c_{\text{Meanpool}}(\text{Train}) = 0.657$, $P < 1 \times 10^{-30}$; $c_{\text{MesoNet}}(\text{Test}) = 0.643$ versus $c_{\text{Meanpool}}(\text{Test}) = 0.649$, $P = 5.78 \times 10^{-18}$, t-test) but significantly outperformed Meanpool on the TCGA dataset, for which the performance of Meanpool was much lower ($c_{\text{MesoNet}}(\text{TCGA}) = 0.656$ versus $c_{\text{Meanpool}}(\text{TCGA}) = 0.581$, $P = 2.66 \times 10^{-20}$, t-test), showing that MesoNet is robust when applied to an independent dataset (Fig. 2a), in contrast to aggregated methods such as Meanpool, in accordance with previous studies[22]. MesoNet still outperformed classical models integrating age and gender (Extended Fig. 2). The TCGA dataset was graded following the 2015 WHO criteria[26], similarly to the MESOBANK dataset (Supplementary Table 2).

Patients diagnosed with SMM are known to be associated with a worse prognosis[8]. In the Test dataset, this subgroup of 60 patients had a median survival of 7.2 months. By comparison, the set of 60 patients predicted to have the worst prognosis by MesoNet consisted of a mix of the three histological types, with only 34% of SMM-classified patients and 40% of EMM-classified patients. These two subgroups had comparable outcomes ($P = 0.53$, Fig. 2b, log-rank test), showing that MesoNet can extract predictive features that transcend the current histological classification and can identify a subgroup of epithelioid patients with a very poor prognosis.

Grade 1 EMM patients are considered to have the best prognosis[6]. This group of 80 patients in the Test dataset had a median survival of 28.1 months. The 80 patients predicted to have the best survival by MesoNet were all epithelioid patients, with similar prognoses, but with a mix of different grades, showing that MesoNet can also extract predictive features orthogonal to tumor grade (Fig. 2c, $P = 0.3$, log-rank test).

The main advantage of MesoNet over histopathology-based classification is its ability to output a continuous risk score, instead of a discrete patient categorization. We split patients from the test set into equivalent groups of low, intermediate and high risk to provide a fair comparison. As expected, these three groups had significantly different outcomes over the entire dataset (Fig. 2d, $P = 2.86 \times 10^{-18}$, log-rank test). More importantly, we were also able to identify similar subgroups with significantly different outcomes amongst the set of EMM patients (Fig. 2d, $P = 3.03 \times 10^{-8}$, log-rank test) and within each grade (EMM grade 1, $P = 0.016$; EMM grade 2, $P = 0.005$; EMM grade 3, $P = 0.022$, log-rank test). This shows that MesoNet can provide risk scores independently of histological type and grade.

Another important aspect of MesoNet is its interpretability. The tile scoring system allows us to score each tile for a given patient. The resulting distribution is shifted towards negative values (resp. positive) for patients with a good (resp. poor) prognosis (Fig. 3a). This scoring system could provide a companion tool for pathologists to identify known or new regions of interest associated with survival for each patient. We assessed the robustness of MesoNet to tissue sampling by comparing the similarity of predictive scores obtained for multiple biopsies from the same patient to the predictive score from biopsies of other patients (see 'Robustness of the prediction to tissue sampling' section of the Extended Methods). The intra-individual variability was significantly lower than the inter-individual variability (two-sided t-test; $P = 9.66 \times 10^{-20}$), indicating that MesoNet is robust to tissue sampling (Extended Fig. 3).

We next sought to obtain an overall understanding of histological predictive features by aggregating all the tile scores from the 2,981 slides of MESOBANK into a single distribution and extracting the ones associated with high and low survival across patients (Fig. 3b). Two expert pathologists manually and independently reviewed these extremal tiles ($n = 42$, Extended Fig. 4a). Most low survival associated tiles were mainly localized in stromal regions, instead of within the tumors (ratio(low (resp high) survival tiles in stroma) = 34/42 (resp. 2/42), Fisher test, $P = 2.8 \times 10^{-13}$, Extended Fig. 4b). These tiles were moreover associated with a transitional pattern, a provisional histopathological pattern recently described[2] (ratio(transitional pattern in low (resp. high) survival tiles) = 11/42 (resp. 0/42), Fisher test, $P = 4.6 \times 10^{-4}$, Extended Fig. 4b), corresponding to a higher grade cellular morphology (ratio(grade 3 in low (resp. high) survival tiles) = 31/42 (resp. 0/42), Fisher test, $P = 9.1 \times 10^{-14}$, Extended Fig. 4b) with atypical nuclei, and a characteristic stromal response, consisting of cancer-associated fibroblasts with small vessels unevenly distributed together with inflammatory cells (Extended Fig. 4c). Other low survival tiles were focused on areas of vacuolated and atypical cells in a dense collagenous stromal response, for which the malignancy of the region cannot be confidently assessed by the pathologist during an initial assessment without immunohistochemistry (ratio(vacuolization in low (resp. high) survival tiles) = 27/42 (resp. 0/42), Fisher test, $P = 2.6 \times 10^{-11}$, Extended Fig. 4b). We named these tiles 'tiles of unknown significance' (TUS, Extended Fig. 4d). In contrast, high survival tiles showed a tubular architecture and were well vascularized (Extended Fig. 4b).

The two pathologists also analyzed the specific histological features differentiating predictive and non-predictive tiles for low and high survival tiles, separately. For the most predictive tiles ($n = 21$ by subgroup), we extracted nine predictive and non-predictive tiles based on the tile scores that were the most similar according to the vector of coefficient (Extended Fig. 5). Conversely, predictive high survival tiles were of lower-grade tumors, less pleomorphic, atypical and showed a greater inflammatory response (Extended Fig. 6). Low survival tiles were of higher-grade tumors, more pleomorphic, atypical and showing a lower inflammatory response (Extended Fig. 6). Moreover, independent of the patient classification, the overall presence of sarcomatoid patterns within a WSI was associated with a worse prognosis, independent of the pathologist diagnosis. These results validate that high pleomorphism, atypia and a lower inflammatory response are consistently associated with a lower survival and should be taken into account by pathologists for MM diagnosis and staging.

We then sought to analyze low survival tiles in the set of EMM patients with poor prognosis. We first confirmed that a large proportion of poor prognosis tiles are associated with a sarcomatoid component (Fig. 4a, 39% of cases), showing that MesoNet can assist in detecting sarcomatoid regions that may have been missed during the initial diagnosis, as well as the importance of these regions for disease outcome. Second, we found that a large proportion of poor prognosis tiles were of the epithelioid component (Fig. 4a, 39% of cases), transitional or 'unknown' components (Fig. 4a, 22% of cases). Interestingly, recent studies demonstrated that a pleomorphic pattern was predictive of tumor aggressiveness in patients with EMM[9,27], in accordance with our findings. Our analysis also showed that the poor survival tiles were mainly located in stromal regions (Fig. 4b), validating the importance of the tumor microenvironment (TME)[28]. As expected, the tiles associated with high survival were all associated with an epithelioid component (data not shown).

CDKN2A/p16 expression and homozygous deletion (HD) were available for a subset of the MESOBANK (Supplementary Table 1) samples. While *p16* expression and HD were associated with a worse prognosis (data not shown) in accordance with previous studies[10,29], MesoNet predictive performance was significantly better than a predictive model based on *p16* expression/homozygous deletion only (c_{MesoNet}(*p16* measured patients in test) = 0.726 versus

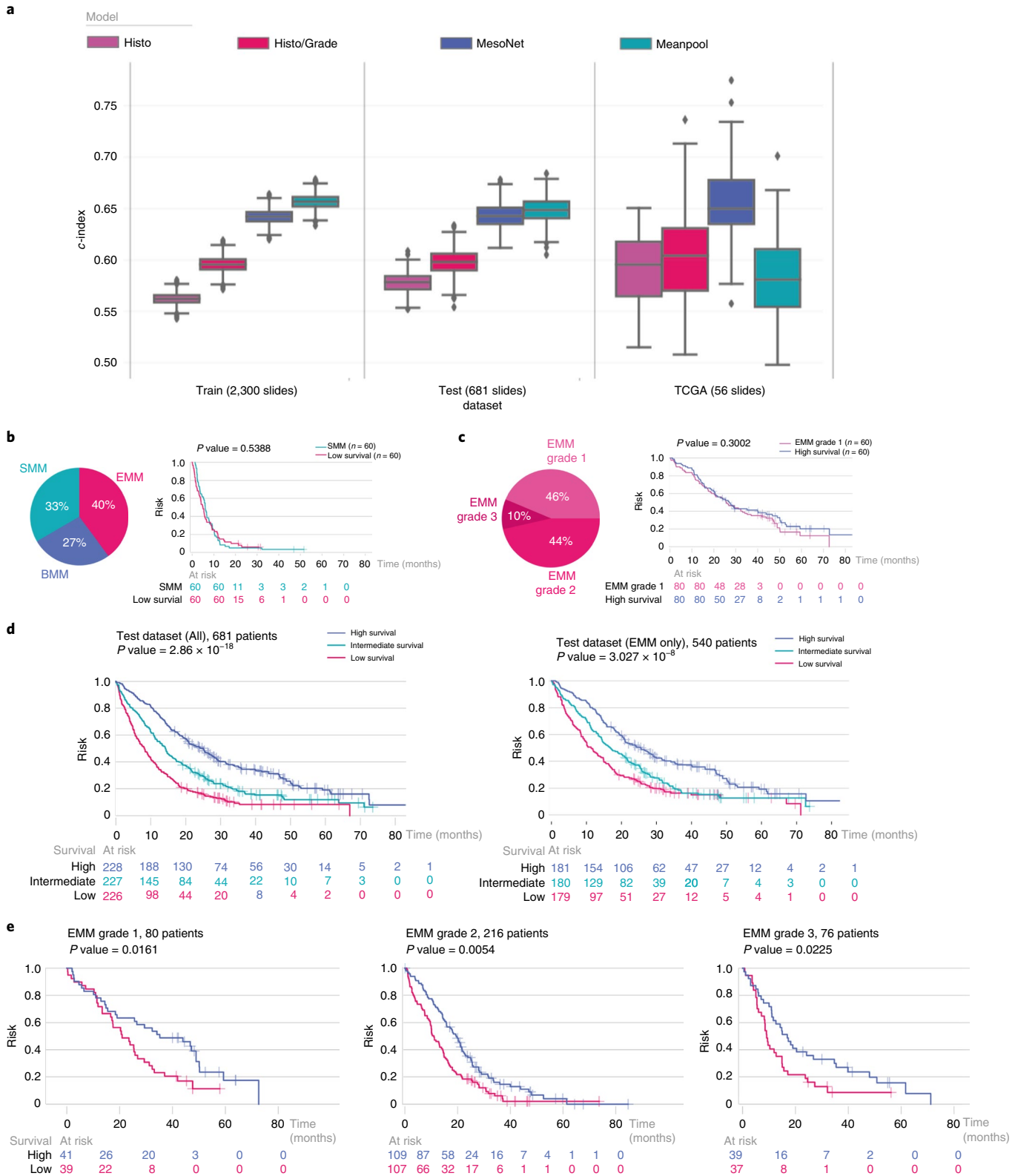


Fig. 2 | Comparison of the performance between MesoNet and other histology-based models to predict MM patient overall survival. a, Distribution of the *c* index for the different predictive models on the Train ($n = 2,300$ slides), Test ($n = 681$ slides) and TCGA ($n = 56$ slides) datasets. **b**, Histological repartition of the 60 patients from the Test dataset with the worst prognosis, as predicted by MesoNet and a comparison of the outcome for the 60 sarcomatoid patients. **c**, Grade distribution of the 80 patients from the Test dataset with the best prognosis, as predicted by MesoNet and a comparison with the outcome for the 80 grade 1 epithelioid patients. **d**, Survival analysis of the whole Test dataset (versus the EMM Test dataset) on the left (versus the right), split evenly into three subgroups of high, intermediate and low survival based on the prediction of MesoNet. **e**, Survival analysis of grade 1, 2 and 3 EMM patients, split evenly into two subgroups based on the prediction of MesoNet. The number of patients per EMM grade is indicated at the top of each graph in **e**. A two-sided log-rank test was performed in **b**–**e** and the exact *P* values are depicted in each panel. For the boxplots, whiskers represent the minima and the maxima. The middle line within the box represents the median. The upper and lower boundaries of the whiskers represent the 25th and 75th percentiles, respectively.

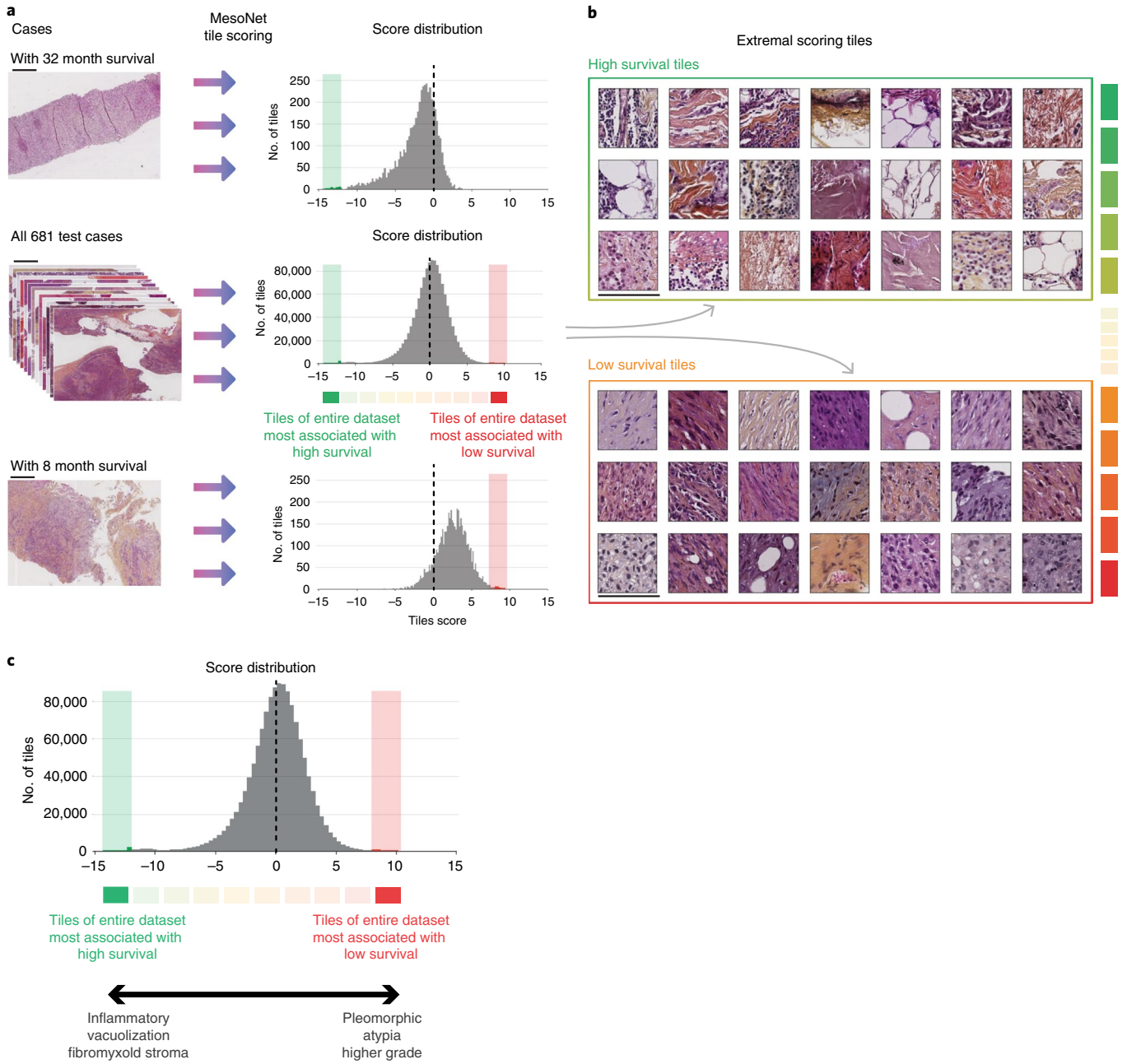


Fig. 3 | Extraction of extremal tiles associated with prognosis to identify regions of interest. a, The distributions of tile scores obtained from a WSI using MesoNet for a patient with a good prognosis is shown at the top and with a bad prognosis at the bottom and aggregated in a single distribution for all patients in the MESOPATH/MESOBANK dataset in the middle ($n = 2{,}981$ cases). Scale bars, 5 mm. **b**, Extremal tiles associated with overall high and low survival for all mesothelioma patients were extracted to be reviewed by a pathologist. Scale bars, 112 µm. **c**, Histological characterization of extremal predictive tiles based on higher scores: tiles associated with lower survival are more associated with pleomorphic, atypia and higher-grade regions. Tiles associated with higher survival are more associated with inflamed, vacuolized and fibrotic patterns.

$c_{p16} = 0.625$, t-test; $P < 1 \times 10^{-30}$; c_{MesoNet}(HD*p16* measured patients in test) = 0.706 versus $c_{\text{HD}p16} = 0.59$, $P < 1 \times 10^{-30}$).

We also compared the MesoNet risk score to additional biological correlates available for the TCGA dataset. Only *Ki67* expression was significantly correlated with MesoNet predicted score (correlation = −0.315; Pearson correlation test, $P = 0.02$; Extended Fig. 7), a strong prognostic factor in MM[30].

In conclusion, we have presented a novel and innovative deep learning framework for prognosis and prediction in MM based on histology slides. Our model, called MesoNet, was trained on the MESOBANK dataset and effectively predicts the OS of mesothelioma patients solely from WSIs of large thoracoscopy or small size needle biopsies. MesoNet is robust to external validation, and the CHOWDER framework[22] is broadly applicable to other tumor and data types or, more importantly, to predict the response to treatment in a clinical setting. In addition, MesoNet can identify known predictive histological features, such as sarcomatoid components, in samples originally classified as epithelioid, which could be useful in assisting pathologists in the histological classification of new biopsies. It can also identify new predictive histological features in regions that were not previously known to contribute to the prognosis, such as stromal regions. Similar findings have already

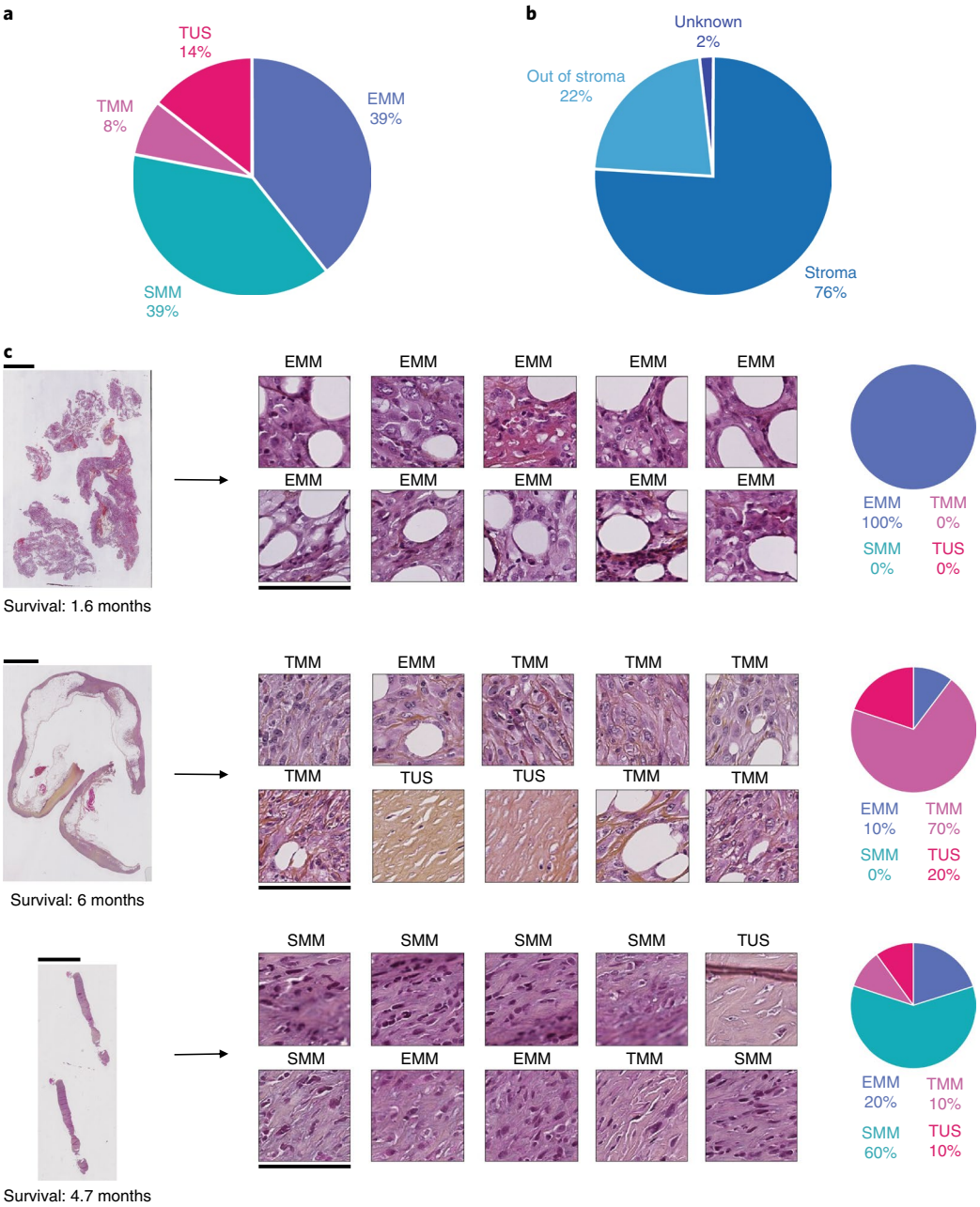


Fig. 4 | Histological characterization and localization of tiles of interest in EMM patients with a poor prognosis. a, Extremal tiles associated with low survival ($n=50$) were mostly composed of epithelioid (EMM) and sarcomatoid (SMM) patterns, but a few were composed of transitional patterns (TMM) and patterns of unknown significance (TUS). **b**, Extremal tiles associated with low survival ($n=50$) were located primarily in stromal regions. **c**, We visualized various scenarios predicted by MesoNet. Patient 1 had an overall survival of 1.6 months, but had tiles consisting of epithelioid components only. Patient 2 had an overall survival of 6 months and had tiles consisting of mainly transitional components. Patient 3 had an overall survival of 4.7 months and had mainly sarcomatoid tiles, which might have been missed at diagnosis. WSI scale bars, 5 mm. Tile scale bars, 112 µm.

been observed in other cancer types[17,31] where a strong influence of the micro environment, and particularly the stroma, is associated with the response to therapies. This focus on the stroma fits with recent analyses of the TME in MM, which showed that ratios of infiltrated macrophages/lymphocytes were important prognostic factors[32]. Altogether, these findings provide the foundation for future investigations into the TME impact on MM to direct target discovery research for a cancer without curative treatment.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41591-019-0583-3.

Received: 13 February 2019; Accepted: 15 August 2019;
Published online: 7 October 2019

References

1. Galateau-Sallé, F., Churg, A., Roggli, V. & Travis, W. D. The 2015 World Health Organization classification of tumors of the pleura: advances since the 2004 classification. *J. Thorac. Oncol.* **11**, 142–154 (2016).
2. Galateau-Sallé, F. et al. New insights on diagnostic reproducibility of biphasic mesotheliomas: a multi-institutional evaluation by the International

Mesothelioma Panel from the MESOPATH reference center. *J. Thorac. Oncol.* **13**, 1189–1203 (2018).
3. Noonan, C. W. Environmental asbestos exposure and mesothelioma. *Ann. Transl. Med.* **5**, 234 (2017).
4. Lacourt, A. et al. Dose–time-response association between occupational asbestos exposure and pleural mesothelioma. *Occup. Environ. Med.* **74**, 691–697 (2017).
5. Robinson, B. W. S. & Lake, R. A. Advances in malignant mesothelioma. *N. Engl. J. Med.* **353**, 1591–1603 (2005).
6. Yap, T. A., Aerts, J. G., Popat, S. & Fennell, D. A. Novel insights into mesothelioma biology and implications for therapy. *Nat. Rev. Cancer* **17**, 475–488 (2017).
7. Opitz, I. et al. A new prognostic score supporting treatment allocation for multimodality therapy for malignant pleural mesothelioma: a review of 12 years' experience. *J. Thorac. Oncol.* **10**, 1634–1641 (2015).
8. Kindler, H. L. et al. Treatment of malignant pleural mesothelioma: American Society of Clinical Oncology clinical practice guideline. *J. Clin. Oncol.* **36**, 1343–1373 (2018).
9. Brcic, L., Vlacic, G., Quehenberger, F. & Kern, I. Reproducibility of malignant pleural mesothelioma histopathologic subtyping. *Arch. Pathol. Lab. Med.* **142**, 747–752 (2018).
10. Hmeljak, J. et al. Integrative molecular characterization of malignant pleural mesothelioma. *Cancer Discov.* **8**, 1548–1565 (2018).
11. Shrestha, R. et al. BAP1 haploinsufficiency predicts a distinct immunogenic class of malignant peritoneal mesothelioma. *Genom. Med.* **11**, 8 (2019).
12. Yu, K. H. et al. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nat. Commun.* **7**, 12474 (2016).
13. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **25**, 1090–1098 (2012).
14. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
15. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
16. Hou, L. et al. Patch-based convolutional neural network for whole slide tissue image classification. In *Proceedings of 2016 IEEE Conference Computer Vision and Pattern Recognitition* (IEEE, 2016); https://doi.org/10.1109/CVPR.2016.266
17. Coudray, N. et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat. Med.* **24**, 1559–1567 (2018).
18. Mobadersany, P. et al. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc. Natl Acad. Sci. USA* **115**, 2970–2979 (2018).
19. Campanella, G. et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.* **25**, 1301–1309 (2019).
20. Schaumberg, A. J. et al. Large-scale annotation of histopathology images from social media. Preprint at https://doi.org/10.1101/396663 (2018).
21. Nagpal, K. et al. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. *Npj Digit. Med.* **2**, 48 (2019).
22. Courtiol, P., Tramel, E. W., Sanselme, M. & Wainrib, G. Classification and disease localization in histopathology using only global labels: a weakly-supervised approach. Preprint at https://arxiv.org/abs/1802.02212 (2018).
23. Zarella, M. D. et al. A practical guide to whole slide imaging. *Arch. Pathol. Lab. Med.* **143**, 222–234 (2019).
24. Mukhopadhyay, S. et al. Whole slide imaging versus microscopy for primary diagnosis in surgical pathology. *Am. J. Surg. Pathol.* **42**, 1 (2018).
25. Galateau-sallé, F. et al. [The French mesothelioma network from 1998 to 2013]. *Ann. Pathol. Elsevier Masson* 34, 51–63 (2014).
26. Baas, P. et al. Malignant pleural mesothelioma: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* **21**, 126–169 (2015).
27. Kadota, K. et al. Pleomorphic epithelioid diffuse malignant pleural mesothelioma: a clinicopathological review and conceptual proposal to reclassify as biphasic or sarcomatoid mesothelioma. *J. Thorac. Oncol.* **6**, 896–904 (2011).
28. Junttila, M. R. & De Sauvage, F. J. Influence of tumour micro-environment heterogeneity on therapeutic response. *Nature* **501**, 346–354 (2013).
29. Dacic, S. et al. Prognostic significance of p16/cdkn2a loss in pleural malignant mesotheliomas. *Virchows Arch.* **453**, 627–635 (2008).
30. Pillai, K., Pourgholami, M. H., Chua, T. C. & Morris, D. L. Prognostic significance of Ki67 expression in malignant peritoneal mesothelioma. *Am. J. Clin. Oncol. Cancer Clin. Trials* **38**, 388–394 (2015).
31. Beck, A. H. et al. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Sci. Transl. Med.* **3**, 108ra113 (2011).
32. Ujiie, H. et al. The tumoral and stromal immune microenvironment in malignant pleural mesothelioma: a comprehensive analysis reveals prognostic immune markers. *Oncoimmunology* **4**, 1–9 (2015).

Acknowledgements

We thank the Owkin Lab for interesting discussions and the Centre Léon Bérard for providing access to the MESOBANK/MESOPATH dataset. We thank the MESOPATH consortium—I. Abd-Alsamad, H. Begueret, E. Brambilla, F. Capron, M.-C. Copin, C. Danel, A.-Y. de Lajartre, E. Foulet-Roge, F. Galateau-Sallé, L. Garbe, O. Groussard, S. Giusiano, V. Hofman, S. Lantejuoul, J.-M. Piquenot, I. Rouquette, S. Sagan, F. Thivolet-Bejui and J.-M. Vignaud—for the histological typing of the mesothelioma samples. We also thank the TCGA network for providing the external validation set (http://cancergenome.nih.gov/).

Author contributions

P.C., C.M., M.M., E.P., S.P., M.S., P.M., S.T., M.Z., N.L.S., N.G., O.E., A.G.N., J.-Y.B., F.G.-S., G.W. and T.C. conceived the idea for this paper. P.C., C.M. and M.M. implemented the analysis. A.G.N. and F.G.-S. reviewed the tiles of interest. P.C., C.M., M.M., E.P., S.P., M.S., P.M., S.T., M.Z., N.L.S., N.G., O.E., A.G.N., J.-Y.B., F.G.-S., G.W. and T.C. contributed to the writing of the manuscript.

Competing interests

The authors declare the following competing interests. Employment: P.C., C.M., M.M., E.P., S.P., M.S., P.M., S.T., M.Z., T.C. and G.W. are employed by Owkin. Advisory: O.E. is on the scientific advisory board of Owkin and equity holder in the company. Financial: Owkin has submitted a patent application entitled 'Systems and methods for mesothelioma feature detection and enhanced prognosis or response to treatment' listing J.-Y.B., F.G.-S., T.C., G.W., P.C. and C.M. as the co-inventors.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41591-019-0583-3.

Supplementary information is available for this paper at https://doi.org/10.1038/s41591-019-0583-3.

Correspondence and requests for materials should be addressed to T.C.

Peer Review Information Javier Carmona was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Methods

Datasets description. The MESOPATH/MESOBANK database is an exhaustive repository of French clinical data and histological samples pertaining to mesothelioma. Each French suspected case of mesothelioma is registered in this database and has been histologically certified through a standardized procedure and evaluated by three pathologist experts in the field without knowledge of prior diagnosis, exposure or outcome. The three pathologists systematically reviewed slides stained by hematoxylin, eosin and saffron (HES), and by a panel of 10 antibodies as follows: calretinin, EMA, CK5/6, pancytokeratin (AE1/AE3), WT1 and mesothelin (positive markers), CEA moabs, BerEP4, TTF-1, ER alpha, PAX8 and p40 (negative markers). The representative sections for evaluation were systematically scanned using a Leica AT2-400 scanner (Leica) at ×40 magnification, associated with the Teleslide (tribvn) to obtain.scn files, which were then archived at the Léon Bérard Cancer Center with their epidemiological and clinical biological annotations. Pathology diagnosis were made following the WHO recommendation[1], with classification into one of three types: epithelioid, sarcomatoid or biphasic. Patients were classified as biphasic if the sample contained at least 10% of both epithelioid and sarcomatoid components.

This cohort consisted of 2,981 patients for which digitized HES-stained mesothelioma histology slides (latter named WSIs), as well as additional clinical information, were available (Supplementary Table 1). A subset of the samples were collected through needle biopsies ($n=38$). All the others were thoracoscopy biopsies.

For our validation set, we also used the diagnostic images from 56 mesothelioma patients from the TCGA (one slide per patient). This dataset was used because TCGA slides were collected at different centers from the MESOBANK slides, using different slide coloration (H&E for TCGA versus HES for MESOPATH/MESOBANK) and different patients staging criteria (from ref. [33] for TCGA and from ref. [26] for MESOPATH/MESOBANK). To compare predictive models, TCGA slides were graded following the same criteria as the MESOPATH/MESOBANK dataset (Supplementary Table 2).

Molecular data. *p16* HD status or *p16* expression determined by immunohistochemistry were defined by MESOBANK as previously described[2].

Data partitioning. We randomly partitioned the MESOPATH/MESOBANK dataset into a training dataset of 2,300 slides from 2,300 patients, which was used to train the model, and a test dataset of 681 slides from 681 patients, which was kept entirely separate to assess the performance of each model in an unbiased manner.

Detailed model structure. Our model architecture is divided into five parts (Extended Fig. 1):

- Matter extraction. We first detect the part of the image that indeed contains matter. This segmentation is performed using a U-Net neural-network approach[34]. All pixels are separated between two classes: pixels of the foreground (containing the matter) and background.
- Tiling. We divide the part of the slides for which matter is detected into smaller images, called 'tiles', of fixed size (224×224 pixels). At least 20% of the tile must have been detected as foreground by the U-Net model to be considered as a tile of matter. The number of tiles depends on the size of the matter detected and can vary from a few hundred to 50,000. We thus decided to limit the extraction to 10,000 tiles, taking into consideration the computation time and required memory.
- Feature extraction. This is performed using the state-of-the-art ResNet-50[35]. Already pre-trained for any image recognition task, this network allows us to obtain 2,048 relevant features from each tile. We therefore obtain a matrix of 10,000 (tiles)×2,048 (features) for each slide. If there is not sufficient matter in the slide to extract 10,000 tiles, we perform a zero padding to fill the matrix. At the end of this step, we obtain a tensor with the following dimension: number of slides×10,000×2,048.
- Top and negative instances. We then use a convolutional one-dimensional (1D) layer to create a score for each tile. This convolutional layer performs a weighted sum between all 2,048 features of the tile to obtain this score (weights of this sum are learned by the model). As our convolutional 1D layer is unbiased, we can be certain that all our zero-padding tiles have a score of 0 and thus a reference for a totally uninformative tile. As previously described[22], we then pick the 10 highest and the 10 lowest scores and use them as input for our last step. This architecture ensures which tiles are used to make the predictions and, therefore, how our algorithm predicts the result.
- Multi-layer perceptron (MLP) classifier. The last step consists of a MLP with two fully connected layers of 200 and 100 neurons with sigmoid activation. This is the core of the predictive algorithm that transforms the scores from the tiles to a prediction.

Model improvements. We performed several improvements to the structure of the previous MesoNet model[22]:

- U-Net segmentation. In the original presentation of the MesoNet pipeline, segmentation was carried out using the Otsu algorithm, which is a generally used segmentation method that requires no training. However, this type of segmentation was not sufficiently robust and failed on slides containing artefacts, such as ink marks, for which it returned an empty mast (data not shown). We therefore decided to use the U-Net method. This was trained on a few hundred thumbnails of histology images in which matter was manually selected. Each thumbnail is an image of 256×256 pixels.
- Survival loss. MesoNet was previously applied to classification problems[36]. The last layer of the architecture was thus changed and switched from 'softmax' activation to linear activation, which better fits the prediction of survival and is similar to a regression problem. Furthermore, this problem is not equivalent to a regression, because of censored data, and requires a specific loss function, such as the Cox loss function, which allows the use of information from censored data.
- Auto-encoding. MesoNet can be subject to over fitting, like many models in machine learning. We solved this recurrent problem in machine learning by reducing the dimension of the input of the prediction part of the network to 512 features instead of 2,048. We used an auto-encoder, which consists of a single hidden-layer architecture (of 512 neurons). This prevents our model from over fitting by finding several singular features in the training dataset and also reduces computation time and required memory. We trained it on 200 tiles randomly selected from each slide (a total of 411,400 tiles). The model converged after three epochs to a mean squared error of 0.0235.

Assessment of performance. We used the *c* index to quantify the concordance between the predicted and true survival time of a set of patients. This metric is commonly used to evaluate and compare predictive models with censored data[37]. The *c* index evaluates whether predicted survival times are ranked in the same order as their true survival time. Because of censored data, the actual ordering of patient survival is not always available. For example, if patient i died at t_i and patient j is still alive at t_i and $t_i > t_j$, the actual order of this pair of patients' survival is not available as patient j may or may not have died before t_i. This pair of patients is defined as a non-admissible pair, in opposition to admissible pairs, where the actual order of patient survival is available.

The *c* index is defined as follows:

$$\text{c index} = \text{number of concordant pairs}/\text{number of admissible pairs}$$

where concordant pairs are the pairs of patients that are correctly classified and admissible pairs are the pairs of patients that can be ordered. For example, (i, j) is admissible if patients i and j are not censored, or if patient i dies at $t=k$ and patient j is censored at $t>k$. On the contrary, if patient i died at $t=k'$ and patient j is censored at $t<k'$, (i, j) is not admissible.

Comparative prognostic models. To compare our MesoNet model to a relevant level of performance, we developed three other methods to predict survival from the whole slide images:

- Histo is only based on the histological type and ranks each patient in three categories: poor prognosis for patients with SMM slides, medium prognosis for patient with BMM slides and good prognosis for patient with EMM slides.
- Histo/Grade consists of a XGBoost[38] model trained on six features for each patient (three encoding the histological types (EMM, BMM or SMM) and three encoding the grading information (grade 1, 2 or 3)).
- Mean Pool is a simpler machine learning algorithm that follows the same first three steps as MesoNet (Extended Fig. 1). However, rather than sorting and selecting the tiles, we perform a simple mean aggregation to obtain 512 features representative for our patient. We then train a linear regression to predict survival from those features.

In Extended Fig. 2 we also provide scores for several models where information regarding sex and age is added:

- Sex/age consists of a linear model trained on two features for each patient (one for age and one for sex).
- Histo/Grade/Sex/Age consists of an XGBoost model trained on eight features for each patient (three encoding histological types, three encoding grading information, one for age and one for sex).
- MesoNet/Sex/Age consists of a multimodal model using WSIs and the sex and age of each patient. The model is a variant of MesoNet where in step 5 of the CHOWDER algorithm (presented in Extended Fig. 1) we had a variable for sex and a variable for age along with the 10 largest and 10 lowest scores extracted from the WSI. An MLP is then trained to predict overall survival from those 22 variables.
- Meanpool/Sex/Age consists of a multimodal model using WSIs and the sex and age of each patient. The model is a variant of the Mean Pool model where we had a variable for sex and a variable for age along with the 512 features representative of the WSI. We then train a linear regression to predict survival from those 514 variables.

Robustness of the prediction to tissue sampling. Spatial intratumor heterogeneity might affect the prediction and the interpretation made by MesoNet. Multiple

biopsies (range 2–6) were available for 167 patients in the Test set. We computed the predicted risk scores for each biopsy sample and compared the maximum intra-individual variability to the inter-individual variability (Extended Fig. 6). The intra-individual variability distribution was significantly lower than the inter-individual variability (two-sided *t*-test; $P = 9.66 \times 10^{-20}$), indicating that MesoNet is robust to intratumor heterogeneity.

Statistical methods. A two-sided Student *t*-test was performed to compare the performances between the different predictive models. A log-rank test was performed to compare the significance of different survival curves. A Pearson correlation test was performed to assess the significance of correlation between two covariates. For *P* values lower than the machine precision level (1×10^{-30}), the exact *P* value was not reported but was indicated to be inferior to 1×10^{-30}.

Ethical compliance. All MESOBANK participants signed an informed consent through their regional centers. The dataset was accessed under the approval of the Centre Léon Bérard IRB. CNIL authorization number, 913346; sample collection authorization, CCTRS/CPP/CNIL DR-2014-068; sample collection and usage authorization, DC-2008-586 and AC-2013-1806; substantial modification in 2018, AC-2018-3221; Rrenewal, AC-2019-3426; import–export authorization, IE-2016-858; MR004 declaration number, 2211136.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

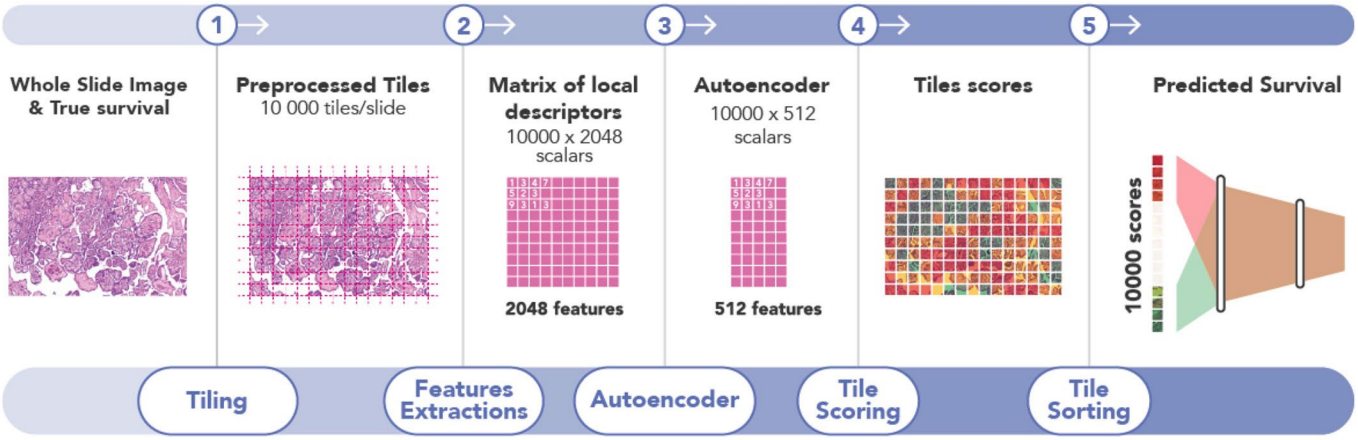
The external validation TCGA dataset is publicly available at the TCGA portal (https://portal.gdc.cancer.gov). We provide a manifest linking to the sample IDs considered in the study (Supplementary Table 4). The MESOBANK/MESOPATH dataset that supports the findings of this study is available from the Centre Léon Bérard, but restrictions apply to its availability (used with permission for the current study), and so it is not publicly available. The data, or a test subset, may be available from the Centre Léon Bérard subject to ethical approvals.

Code availability

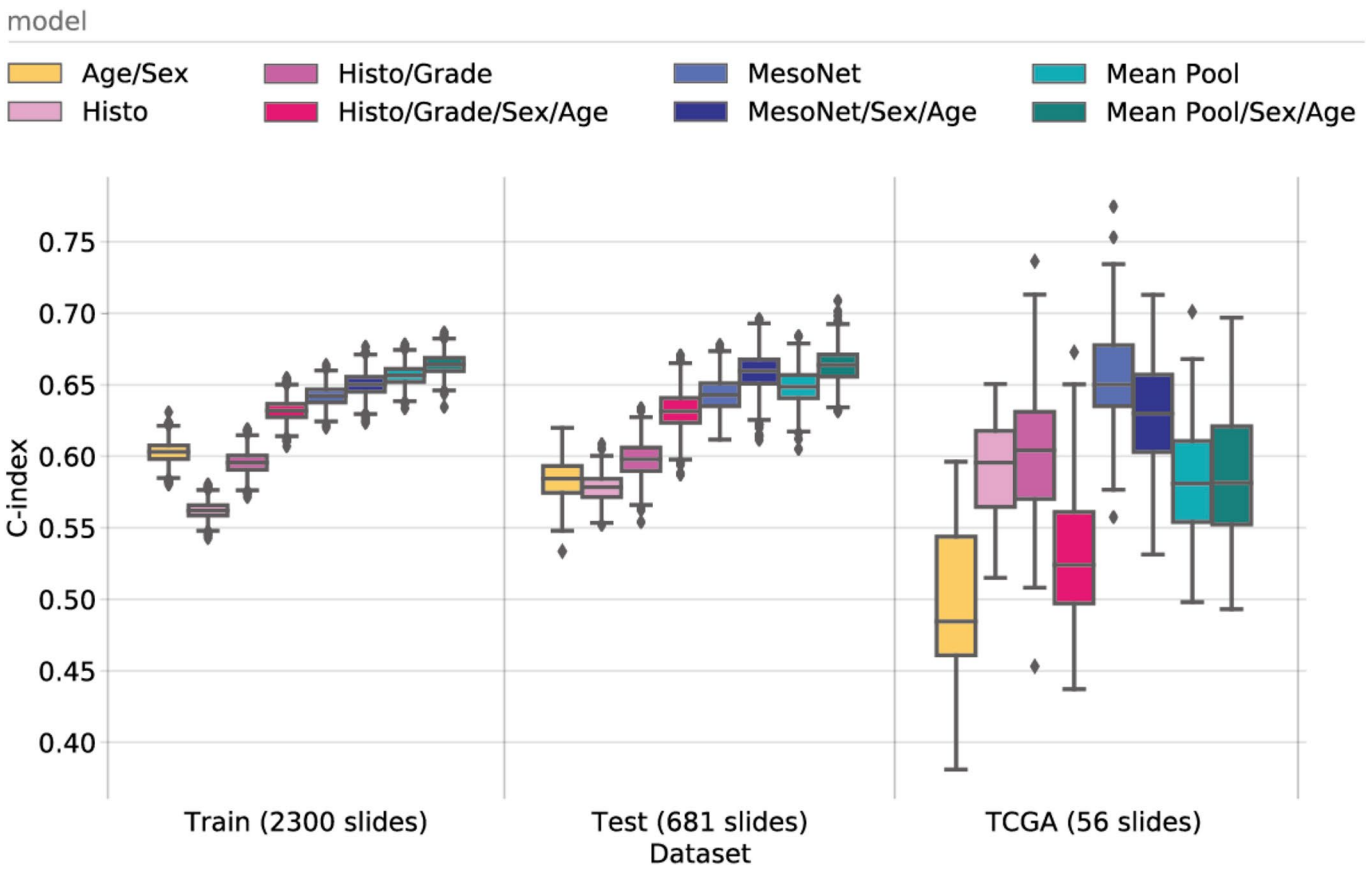
The code used for training the models has a large number of dependencies on internal tooling and its release is therefore not feasible. However, all experiments and implementation details are described thoroughly in the Methods so that it can be independently replicated with non-proprietary libraries.

References

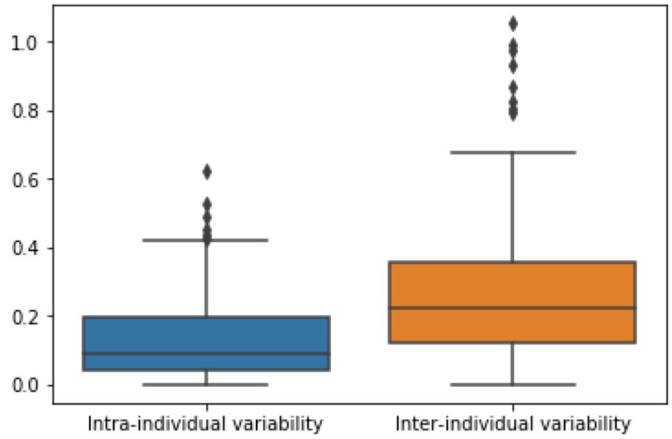
33. Rosen, L. E. et al. Nuclear grade and necrosis predict prognosis in malignant epithelioid pleural mesothelioma: a multi-institutional study. *Mod. Pathol.* **31**, 598–606 (2018).
34. Ronneberger, O., Fischer, P. & Brox, T. U-net: convolutional networks for biomedical image segmentation. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference* 234–241 (Springer, 2015); https://doi.org/10.1007/978-3-319-24574-4_28
35. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2016); https://doi.org/10.1109/CVPR.2016.90
36. Wang, D., Khosla, A., Gargeya, R., Irsha, H. & Beck, A.H. Deep learning for identifying metastatic breast cancer. Preprint at https://arxiv.org/abs/1606.05718 (2016).
37. Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B. & Wei, L. J. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat. Med.* **30**, 1105–1117 (2011).
38. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (ACM Press, 2016); https://doi.org/10.1145/2939672.2939785



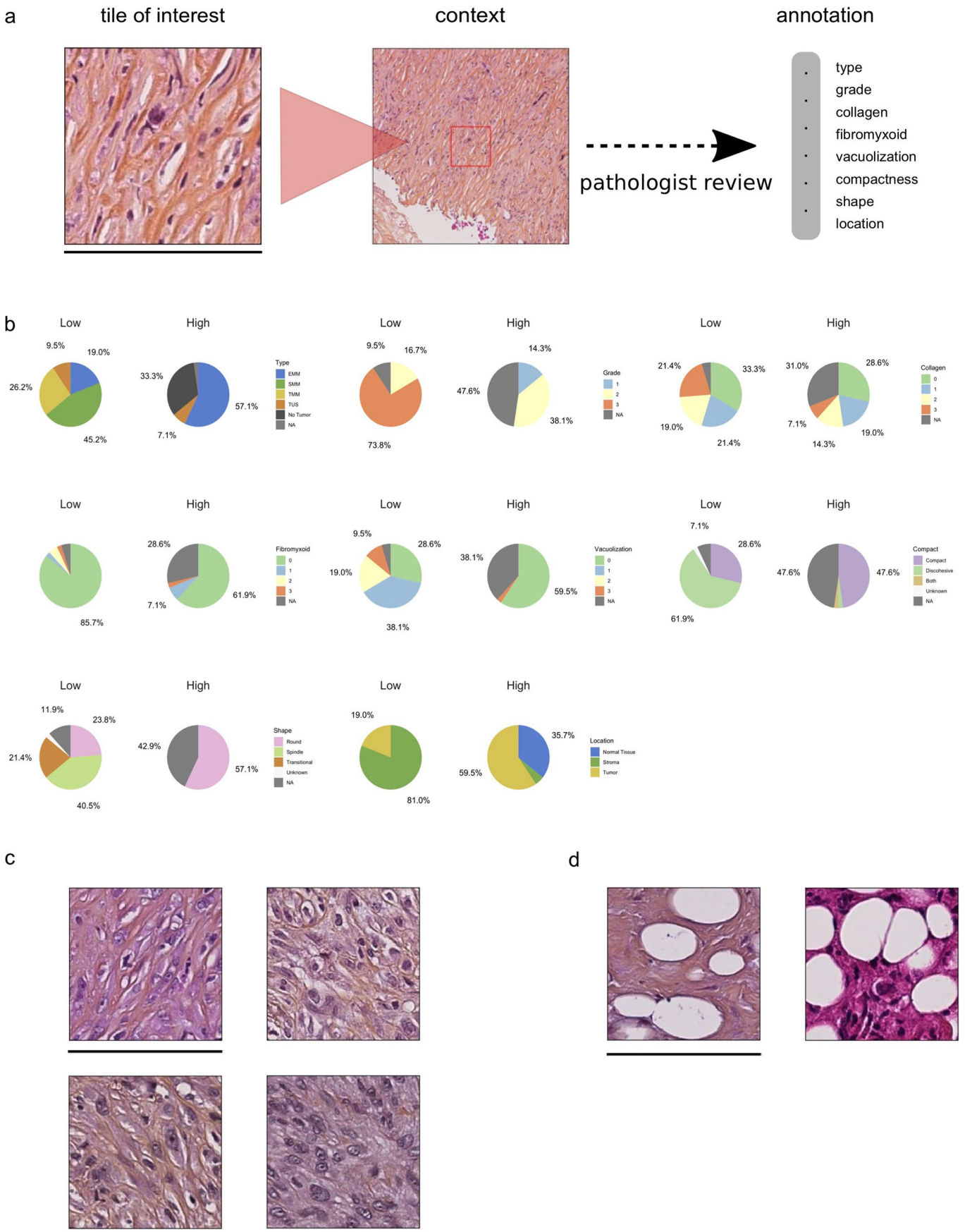
Extended Data Fig. 1 | MesoNet layout. Our prognosis prediction is composed of five steps. (1) The 'tiling' process divides the whole slide image into small tiles of 224 × 224 pixels. (2) Features are then extracted from each tile using ResNet-50. (3) Features are then auto-encoded to reduce the dimension to 512 features per tile. (4) We use a convolutional 1D layer to score each tile. (5) Tiles associated with the largest and lowest scores are then retained to train a multilayer perceptron (MLP) to predict overall survival of the patient.



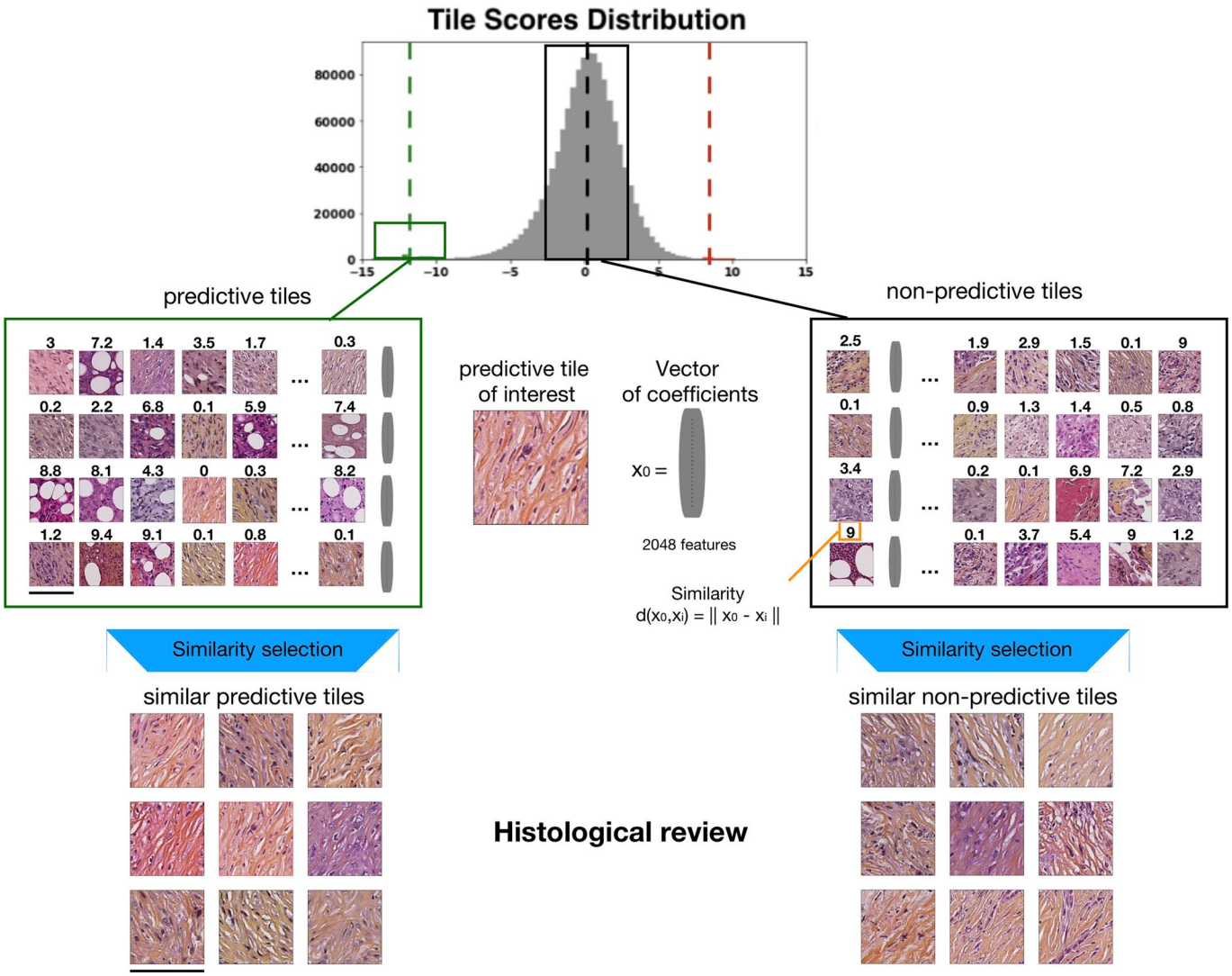
Extended Data Fig. 2 | Comparison of the performance between MesoNet and models including additional non-pathology variables such as age and sex to predict MM patient overall survival. The *c* index distribution for different predictive models. For the box plots, whiskers represent the minima and maxima. The middle line within the box represents the median. The upper and lower boundaries of the whiskers represent the 25th and 75th percentiles, respectively. A two-sided *t*-test was performed to determine significance.



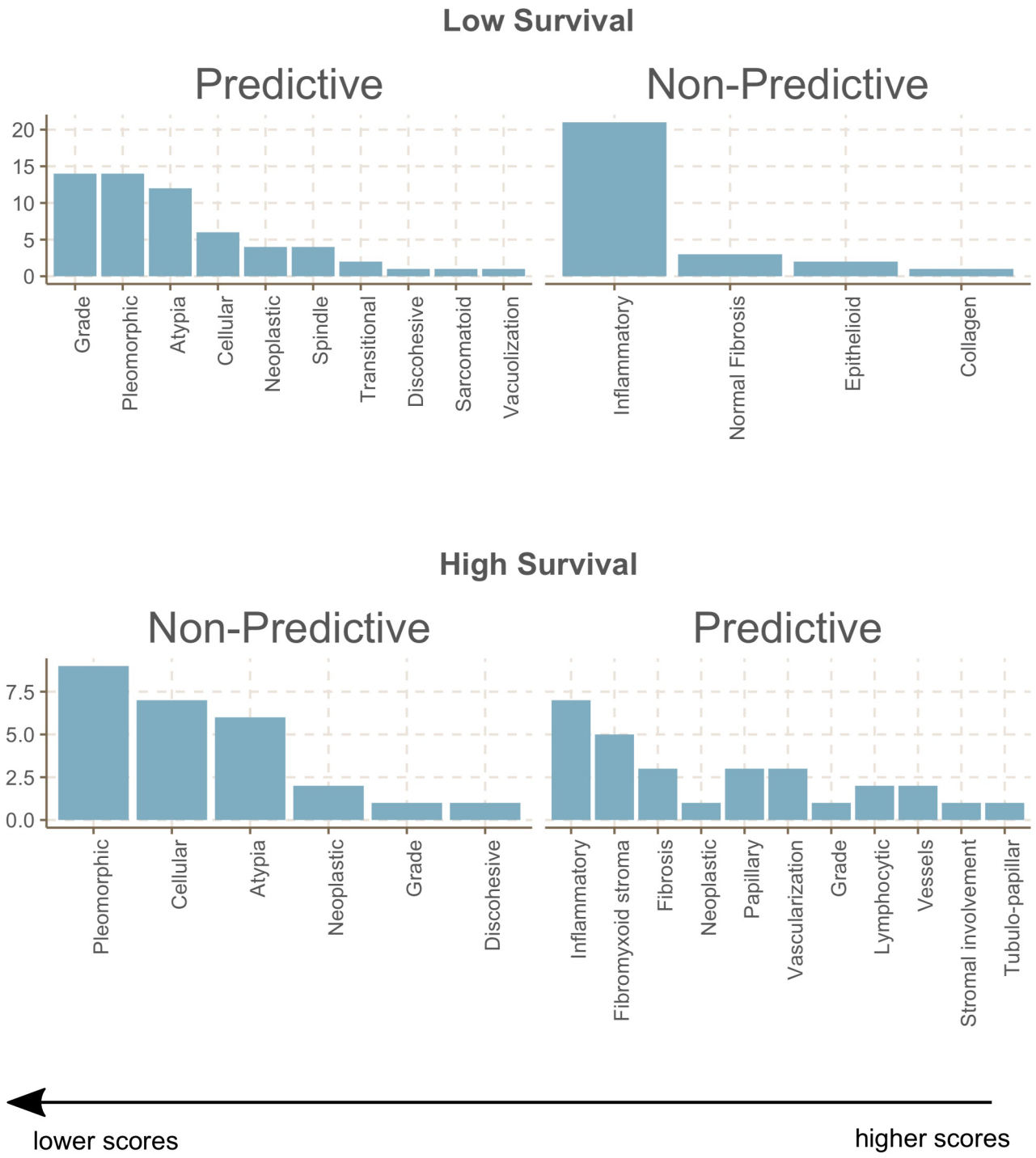
Extended Data Fig. 3 | Robustness of MesoNet to tissue sampling. Distributions of variability of the MesoNet prediction for the 167 patients where multiple biopsies were available. The boxplot on the left side represents the distribution of maximal intra-individual variability, that is, the absolute difference between the minimum score and maximum score predicted for all the slides from a single patient for all patients with multiple biopsies ($n=167$ patients). The boxplot on the right side represents the inter-individual variability, that is, the absolute difference between the mean of the scores for all the slides of a given patient to the mean of the score for all patients with multiple biopsies ($n=167$ patients). For the two boxplots, whiskers represent the minima and maxima. The middle line within the box represents the median. The upper and lower boundaries of the whiskers represent the 25th and 75th percentiles, respectively. Significance was calculated using a two-sided *t*-test ($P=9.66\times10^{-20}$).



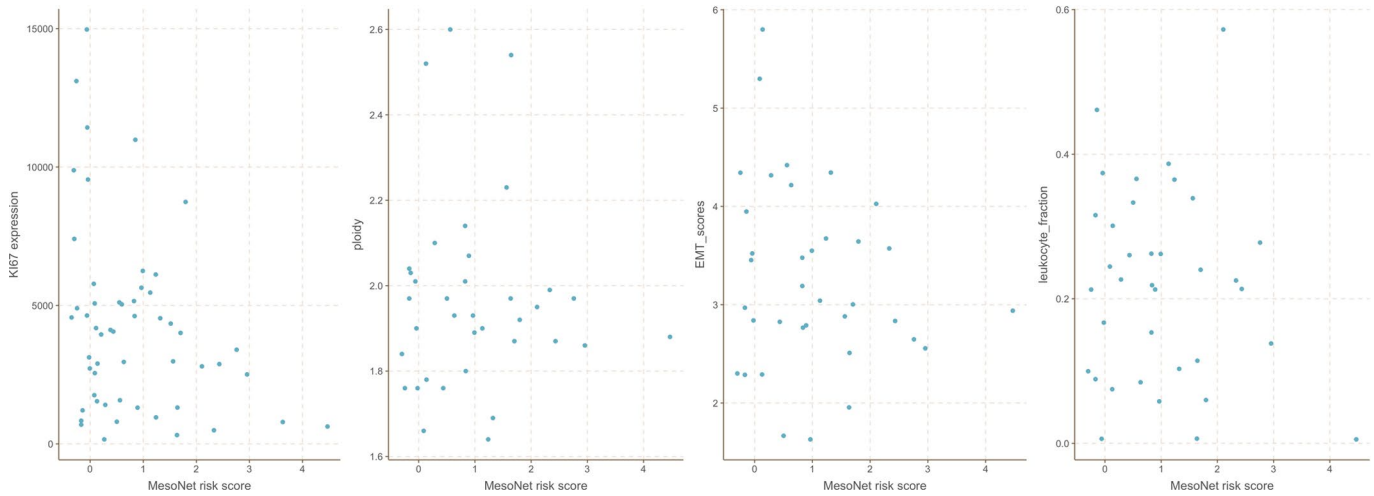
Extended Data Fig. 4 | Histological review of tiles of interest. a, Schematic of the reviewing process. **b**, Repartition of features of interests in low survival tiles ($n = 42$) and high survival tiles ($n = 42$). **c**, Tiles of low survival with a transitional pattern. **d**, Tiles of unknown significance, TUS. Scale bars, 112 µm.



Extended Data Fig. 5 | Illustration of predictive and non-predictive tiles selection. Similarity of predictive and non-predictive tiles to a given predictive tile of interest, calculated based on the vector of coefficients obtained with ResNet-50. Similar predictive and non-predictive tiles are then reviewed manually for each extremal tile by pathologists. Scale bars, 112 µm.



Extended Data Fig. 6 | Comparative histological analysis of predictive and non-predictive tiles. Histogram of histological features associated with either predictive or non-predictive tiles that are similar to extremal tiles of high and low survival ($n = 42$). The review was performed independently by two mesothelioma pathologists.



Extended Data Fig. 7 | Biological correlates. Correlation between MesoNet risk score and the *Ki67* expression (n = 54 samples), the ploidy level (n = 36 samples), the EMT score (n = 36 samples) and the leukocyte fraction (n = 36 samples) available for the TCGA dataset.

nature research

Corresponding author(s): Thomas Clozel

Last updated by author(s): Aug 5, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Histological slides were systematically scanned using a LEICA AT2 -400 (Leica) at 40X magnification associated with the Teleslide (tribvn) to obtain .scn files
Data analysis	Jupyter 4.4.0 Python 3.6.4 matplotlib==2.1.2 numpy==1.15.4 tensorflow-gpu==1.10.1 scipy==1.0.0 openslide-python==1.1.1 pandas==0.22.0 scikit-learn==0.19.1 Keras==2.2.4

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The external validation TCGA dataset is publicly available at the TCGA portal (https://portal.gdc.cancer.gov). We added a manifest linking to the sample ids

considered in the study (Extended data table 4). The MESOBANK/MESOPATH dataset that supports the findings of this study is available from the Centre Léon Bérard but restrictions apply to the availability, which were used with permission for the current study, and so are not publicly available. The data , or a test subset, may be available from the Centre Léon Bérard subject to ethical approvals.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The sample size (n=2,981 patients) was obtained because it is the exhaustive repository of mesothelioma patients from the MESOBANK database that was digitized.
Data exclusions	Samples not classified as Epithelioid, Biphasic and Sarcomatoid were removed from the study.
Replication	The results have been generated via a python notebook to guarantee that they can be reproduced easily. Experiments with a randomization part were performed multiple times to assess the reproducibility of the results.
Randomization	Randomization of patients for cross-validation was performed completely at random without any stratification.
Blinding	Histological typing was performed by expert pathologists from the MESOPATH consortium without knowledge of any clinical information from the patients. For the review of regions of interests, pathologists were independently assigned regions of interest to review and were not able to communicate on their results to each other so that there is no bias in each pathologist review.Histological typing was performed by expert pathologists from the MESOPATH consortium without knowledge of any clinical information from the patients. For the review of regions of interests, pathologists were independently assigned regions of interest to review and were not able to communicate on their results to each other so that there is no bias in each pathologist review.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	Samples from all patients diagnosed with mesothelioma in France are centralized into a national repository (MESOBANK / MESOPATH). Access to digitized samples was provided by the Centre Léon Bérard.
Recruitment	All patients diagnosed with mesothelioma in France were recruited in the study. Slides were collected in each center and digitized and stored in an exhaustive repository by the Centre Léon Bérard. Histological typing was performed by 15 pathologists.
Ethics oversight	All participants signed an informed consent through their regional centers. The dataset was accessed under the approval of the Centre Léon Bérard IRB. CNIL authorization number: 913346. Sample Collection authorization CCTRS/CPP/CNIL DR-2014-068. Sample Collection and usage authorization DC-2008-586 and AC-2013-1806. Substantial modification in 2018 (AC-2018-3221). Renewal (AC-2019-3426). Import-Export authorization IE-2016-858. MR004 declaration number 2211136.

Note that full information on the approval of the study protocol must also be provided in the manuscript.