Clinical Radiology 74 (2019) 692–696

Contents lists available at ScienceDirect

Clinical Radiology

journal homepage: www.clinicalradiologyonline.net

Comparison of ^{18}F-FDG avidity at PET of benign and malignant pure ground-glass opacities: a paradox? Part II: artificial neural network integration of the PET/CT characteristics of ground-glass opacities to predict their likelihood of malignancy

J.A. Scott [a], S. McDermott [b], A. Kilcoyne [c], Y. Wang [a], E.F. Halpern [d], J.B. Ackman [b,*]

[a] *MGH Department of Radiology, Division of Nuclear Medicine and Molecular Imaging, 55 Fruit Street, Boston, MA 02114, USA*
[b] *MGH Department of Radiology, Division of Thoracic Imaging and Intervention, Founders House 202, 55 Fruit Street, Boston, MA 02114, USA*
[c] *MGH Department of Radiology, Division of Abdominal Imaging, 55 Fruit Street, Boston, MA 02114, USA*
[d] *MGH Institute for Technology Assessment, 101 Merrimac St, Boston, MA 02114, USA*

ARTICLE INFORMATION

Article history:
Received 3 March 2019
Accepted 26 April 2019

AIM: To assess the ability of artificial neural networks (ANNs) to predict the likelihood of malignancy of pure ground-glass opacities (GGOs), using observations from computed tomography (CT) and 2-[^{18}F]-fluoro-2-deoxy-D-glucose (FDG) positron-emission tomography (PET) images and relevant clinical information.

MATERIALS AND METHODS: One hundred and twenty-five cases of pure GGOs described in a previous article were used to train and evaluate the performance of an ANN to predict the likelihood of malignancy in each of the GGOs. Eighty-five cases selected randomly were used for training the network and the remaining 40 cases for testing. The ANN was constructed from the image data and basic clinical information. The predictions of the ANN were compared with blinded expert estimates of the likelihood of malignancy.

RESULTS: The ANN showed excellent predictive value in estimating the likelihood of malignancy (AUC = 0.98±0.02). Employing the optimal cut-off point from the receiver operating characteristic (ROC) curve, the ANN correctly identified 11/11 malignant lesions (sensitivity 100%) and 27/29 benign lesions (specificity 93.1%). The expert readers found 23 lesions indeterminate and correctly identified 17 lesions as benign.

* Guarantor and correspondent: J. B. Ackman, MGH Department of Radiology, Division of Thoracic Imaging and Intervention, Founders House 202, 55 Fruit Street, Boston, MA 02114, USA. Tel.: +1 617 724 4254; fax: +1 617 724 0046.
E-mail address: jackman@mgh.harvard.edu (J.B. Ackman).

https://doi.org/10.1016/j.crad.2019.04.024

CONCLUSION: ANNs have potential to improve diagnostic certainty in the classification of pure GGOs, based upon their CT appearance, intensity of FDG uptake, and relevant clinical information, and may therefore, be useful to help direct clinical and imaging follow-up.

Introduction

Artificial neural networks (ANNs) are computer software programs that mathematically simulate the parallel architecture of the human brain and have been used to perform an increasing variety of imaging-related tasks.[1–5] ANNs have proven useful in modelling complex relationships among predictor variables that are often obscure to clinical observation alone. Previous examples of ANN-based approaches to image interpretation have included lesion detection and evaluation,[6] extraction of physiological parameters from radiography images,[7] and differential diagnosis.[8,9] More recently ANNs have been applied successfully to the staging of non-small cell lung cancer with combined 2-[^{18}F]-fluoro-2-deoxy-D-glucose (FDG) positron-emission tomography/computed tomography (PET/CT).[10] The goal of the present study was to evaluate the ability of ANNs to synthesise subjective and objective observations concerning the image appearance of pure ground-glass opacities (GGOs) on thin-section (1.5 mm) chest CT and ^{18}FDG PET/CT into predictions of the likelihood of malignancy.

In a previous article, hereby referred to as GGO Part I, new observations were reported regarding the CT characteristics and ^{18}F-FDG avidity of benign versus malignant pure GGOs.[11] As an extension of that study, the hypothesis of the present study is that ANNs can be used to synthesise these various observations concerning the GGO, together with limited basic clinical data, into a single numeric probability of malignancy based upon assessment of their image characteristics on PET/CT at a single point in time.

Materials and methods

Approval for this HIPAA-compliant, retrospective study was obtained from the institutional review board. The requirement for informed consent was waived. The patient population and methods of image evaluation are as described in Part I. Previous studies have shown that the ratio of the maximum standardised uptake value (SUV_{max}) to the standardised uptake value of the liver (SUV_{liver}) shows less variability than does the SUV_{max}[12–14] and thus this corrected value (SUV_{corr}) was employed as a network input. These and other ANN inputs, including patient age, history of lung cancer, the mean diameter of the lesion, and the morphology of the lesion, were obtained as described in Part I. Two of the 127 cases reported in GGO Part I could not be used for this analysis because of diffuse liver abnormalities that precluded a meaningful evaluation of the SUV_{liver}.

ANN construction

The ANN was constructed using a software network simulation program (Neuroshell 2, version 4.2; Ward Systems Group, Frederick, MD, USA). A back-propagation algorithm was employed with a logistic activation function applied to a three-node hidden layer. The network was trained until the mean error of the network failed to decrease after 5,000 consecutive iterations. A single output node indicated the likelihood of malignancy on a continuous scale between 0 and 100%. There were six input nodes including patient age, lesion mean diameter, lesion shape (nodular or non-nodular), SUV_{corr} (as SUV max corrected for mean hepatic SUV), history of lung cancer, and lesion location (whether the lesion was located in an upper lobe or not). These input characteristics were chosen based upon the results of GGO Part I. There were three hidden nodes in a single layer. Training and validation sets were strictly independent of one another. The ANN was constructed by randomly separating the set of 125 cases into 85 training cases and 40 test cases, thus assuring independence of the training and testing data. The networks consisted of the six inputs listed above, three nodes in a single hidden layer, and one output node indicating the predicted likelihood of malignancy between 0 and 1.

ANN data analysis

The ANN data were analysed using a receiver operating characteristic (ROC) curve and by comparing its predictions to those of the two expert readers on a case-by-case basis. The optimal cut-off point was determined from ROC analysis of the ANN predictions to determine potential sensitivity and specificity. The ANN predictions were compared with the clinical predictions of the two board-certified, subspecialty-trained radiologists with 7 and 21 years of experience, blinded to all clinical data, to determine the projected utility of the ANN data. The two expert readers categorised the likelihood of malignancy in each case as “benign,” “indeterminate,” or “malignant,” based upon the appearance of the pure GGO on a single CT study. The final diagnosis was made as described in Part I, using principles of standard clinical practice and supported by the most updated, 2017 Fleischner Society Guidelines[15]: a pure GGO was considered malignant if histopathology was available and showed malignant cells or if the GGO increased in attenuation and size over an interval >6 months; a GGO was considered benign if it resolved, decreased in size, and attenuation over a ≥3 month period in the absence of

chemotherapy, had arisen (≥1cm) within 3 months, was stable for >5 years, or had imaging features evolving in a manner consistent with radiation fibrosis or interstitial lung disease. Any pure GGO that did not meet these criteria was characterised as indeterminate.

All calculations were performed on a computer employing an Intel i7-6850K CPU at 3.6 GHz using 24GB 1066 MHz DRAM, and dual NVidia GTX1080Ti GPUs. Statistical analysis was performed with SPSS (version 24, IBM Corp). Numbers are shown±standard error of the mean.

Results

The ROC curve obtained shows an area under the ROC curve (AUC) of 0.981±0.017 (95% confidence range of 0.948–1). The average ANN prediction in benign cases was 9.7±3.5% and in malignant lesions 79.9±9.6% ($p<0.001$). Youden's statistic was 0.931 and the ROC curve showed an optimal cut-off of 16.9%. Using predictions higher than this value to indicate malignancy, 11/11 malignant lesions were correctly identified by the ANN (100% sensitivity) and 27/29 benign lesions were correctly identified (93.1% specificity). Table 1 shows the predictions of the two expert readers on the 40 test cases as well as the ANN predictions. Neither reader predicted malignancy in the test subset of 40 cases. When both readers interpreted the lesion as benign (17 cases), the ANN predicted an average likelihood of malignancy of 3% (0–12.8%). When the two expert readers differed between benign and indeterminate (11 cases), the network predicted an average likelihood of malignancy of 12.5% (0–60.8%). When both expert readers read the lesion as indeterminate (13 cases), the network predicted an average likelihood of malignancy of 74.9% (10.2–100%). There were no predictive errors among the definitive interpretations by either of the expert readers.

Table 1
Diagnostic prediction by subspecialty-trained radiologist readers and by ANN.

Diagnostic prediction by two radiologists, side-by-side	Final diagnosis of pure GGO	ANN prediction of the likelihood of malignancy (%)
Benign–benign	Benign	0
Benign–benign	Benign	0
Benign–benign	Benign	0
Benign–benign	Benign	0
Benign–benign	Benign	0
Benign–indeterminate	Benign	0
Benign–indeterminate	Benign	0
Benign–benign	Benign	0
Benign–benign	Benign	0
Benign–benign	Benign	0
Benign–benign	Benign	0
Benign–benign	Benign	0
Benign–indeterminate	Benign	0
Benign–benign	Benign	3.2
Benign–indeterminate	Benign	3.9
Benign–benign	Benign	7.9
Benign–benign	Benign	8
Benign–benign	Benign	8
Benign–indeterminate	Benign	9.2
Indeterminate–indeterminate	Benign	10.2
Benign–benign	Benign	10.6
Benign–benign	Benign	11
Benign–indeterminate	Benign	11.1
Benign–indeterminate	Benign	11.9
Benign–benign	Benign	12.8
Benign–indeterminate	Benign	13.8
Benign–indeterminate	Benign	14.3
Benign–indeterminate	Benign	60.8
Indeterminate–indeterminate	Benign	85.4
Indeterminate–indeterminate	Malignant	19.5
Indeterminate–indeterminate	Malignant	28.3
Indeterminate–indeterminate	Malignant	47.5
Indeterminate–indeterminate	Malignant	86.5
Indeterminate–indeterminate	Malignant	97.9
Indeterminate–indeterminate	Malignant	98.9
Indeterminate–indeterminate	Malignant	99.7
Indeterminate–indeterminate	Malignant	100
Indeterminate–indeterminate	Malignant	100
Indeterminate–indeterminate	Malignant	100
Indeterminate–indeterminate	Malignant	100

GGO, ground-glass opacity.

Discussion

A major advantage of ANN-based approaches to lesion diagnosis is their ability to capture often unsuspected and unrecognised interactions between disparate parameters that might be overlooked, if these parameters were considered in isolation. For instance, although lesion size and morphology might each be associated with a particular likelihood of malignancy, a potential interaction between these two parameters might change the predictive value of either. Such often-obscure interactions between input variables are discovered and refined during the ANN training process, which iteratively and progressively optimises a global prediction of the presence or absence of a disease. This may not be unlike the manner in which the human brain analyses images, refining differential considerations through conscious and subconscious iterations until a conclusion is reached.

In this investigation, standard iterative, back-propagation training[16] on a set of clinical data was employed until a pre-defined number of successive iterations (5,000) failed to further reduce the cumulative error on the training cases. This trained network was then tested on a different set of validation cases, one to which the network had not been previously exposed. This method ensured that the network did not simply capture idiosyncratic characteristics of the training data and that it was capable of generalising its "learning" to new data. In other words, without an independent validation set, it is possible to create apparently highly accurate ANNs that capture the fine detail of a single population but do not extrapolate to a different population. Such an ANN would be of little clinical value. In addition to an independent validation set, appropriate network training is critical. It is possible to "over-train" a network (perform an excessive number of iterations), such that it memorises all characteristics, even the idiosyncratic ones, of the training set. This can lead to poor generalisation when the network is applied to new cases. An example might be if a network was being trained to discriminate between voter preferences based upon certain demographic information. Were forearm length unwisely included as an input parameter and the training set included disproportionate numbers of persons with long forearms, who happened be of one political persuasion, the network might conclude that a long forearm is an important indicator of political leaning. On the other hand, it is equally important to avoid under-training the ANN as this leads to poor classifications due to an insufficient opportunity to learn: the ANN fails to become an "expert system." There are no clearly defined rules regarding the proper level of network training. Determination of sufficient network training is therefore largely based upon experience and attention to how the ANN performs in independent validation sets.

Although potentially offering significant predictive accuracy, such network-based diagnostic approaches, being of a purely mathematical nature, have the disadvantage of often lacking a logically explicable "rationale," compared to decision tree or logistic regression models.[17] This relative lack of a justification for a specific result can be disturbing to clinicians seeking a logical explanation for a clinical prediction, particularly when the prediction is not an expected one. Did the network make an unexpected prediction because it found an unsuspected clue to the diagnosis or did it simply make an error based upon limitations of the data upon which it was trained?

The ROC analysis showed excellent performance of the ANN on the test data set (AUC=0.981). The optimal cut-off of the ROC curve produced a sensitivity of 100% (11/11) and a specificity of 93.1% (27/29). The ANN was able to correctly characterise a substantial percentage of pure GGOs as malignant, which the two expert readers could only classify as indeterminate. At a single point in time on a single chest CT (the conditions for assessment by the subspecialty-trained radiologists in Part I of this study), it is impossible for any radiologist, no matter how expert, to determine that a pure GGO is malignant. Some features of pure GGOs, however, can be used for their characterisation as benign and were applied. Without access to comparison studies or the results of serial chest CT follow-up, the radiologists, unlike the ANN, were therefore able solely to characterise the lesions as benign or indeterminate.

The largest error made by the ANN was with regard to a benign pure GGO, for which the likelihood of malignancy was predicted as 85.4%. It is likely that a congruence of factors predisposing to malignancy including relatively advanced age, relatively small lesion size, a nodular appearance, a relatively low SUV, and a history lung cancer all influenced the network prediction.

The results of this study support the prevailing opinion that exclusive reliance on the prediction of an ANN to interpret an imaging finding is not appropriate. This caveat is particularly important when the network is trained on a limited data set, as in the present study. Such networks are most judiciously employed as a "second opinion."

Despite the efficacy of machine learning methods in image-based diagnosis, the methods are not without limitations.[18] Their performance is highly dependent upon the training set. In other words, the training set should include a comprehensive representation of the range of manifestations with which the pathology in question might present. Although the study suggests the potential utility of this method, the present cases were insufficient in number to sample the universe of pure GGOs adequately, something that would be required for the general clinical use of such an algorithm. It is also important to bear in mind that ANNs are mathematical structures whose processing of the input data can be resistant to logical explanation. This "black box" nature of the ANN can make an unexpected prediction difficult to interpret.

Limitations

Although general limitations of the study were described in a preceding study,[11] a particular limitation with respect to the ANN analysis was the relatively larger number of benign,

compared to malignant, cases. It is likely that an ANN analysis employing a larger number of patients, with a larger sampling of malignant lesions, could capture more complex inter-relationships between the inputs and permit the inclusion of additional input variables that might further increase predictive accuracy. The lower percentage of malignant GGOs than benign GGOs in the present study, which included "all-comers" in 2011 fulfilling study criteria, reflects the reality that benign GGOs were simply more common than malignant GGOs, at least in 2011, but most likely, in general. In addition, the ANN incorporated clinical information including age and history of lung cancer into its analysis and characterisation of the pure GGO, whereas the radiologists were blinded to clinical information and therefore did not incorporate it into their evaluation of the pure GGO. However, the aim of the present study was to investigate the potential for ANNs to serve as an adjunct, rather than as a replacement, for expert image interpretation.

In conclusion, the results of this investigation suggest that ANNs have the potential to assist the radiologist in estimating the likelihood of malignancy of GGOs on a single ^{18}FDG PET/CT study, based upon their appearance and relevant clinical information.

Conflict of interest

The authors declare no conflict of interest.

References

1. Jiang J, Trundle P, Ren J. Medical image analysis with artificial neural networks. *Comput Med Imaging Graph* 2010;**34**:617–31.
2. Petrick N, Sahiner B, Armato 3rd SG, *et al*. Evaluation of computer-aided detection and diagnosis systems. *Med Phys* 2013;**40**:087001.
3. Shiraishi J, Li Q, Appelbaum D, *et al*. Computer-aided diagnosis and artificial intelligence in clinical imaging. *Semin Nucl Med* 2011;**41**:449–62.
4. Swietlik D, Bandurski T, Lass P. Artificial neural networks in nuclear medicine. *Nucl Med Rev Cent East Eur* 2004;**7**:59–67.
5. Nam JG, Park S, Hwang EJ, *et al*. Development and validation of deep learning-based automatic detection algorithm for malignant pulmonary nodules on chest radiographs. *Radiology* 2019;**290**:218–28.
6. Chen H, Xu Y, Ma Y, *et al*. Neural network ensemble-based computer-aided diagnosis for differentiation of lung nodules on CT images: clinical evaluation. *Acad Radiol* 2010;**17**:595–602.
7. Scott JA. Pulmonary perfusion patterns and pulmonary arterial pressure. *Radiology* 2002;**224**:513–8.
8. Tourassi GD, Floyd CE, Sostman HD, *et al*. Artificial neural network for diagnosis of acute pulmonary embolism: effect of case and observer selection. *Radiology* 1995;**194**:889–93.
9. Wang S, Summers RM. Machine learning and radiology. *Med Image Anal* 2012;**16**:933–51.
10. Toney LK, Vesselle HJ. Neural networks for nodal staging of non-small cell lung cancer with FDG PET and CT: importance of combining uptake values and sizes of nodes and primary tumor. *Radiology* 2014;**270**:91–8.
11. McDermott S, Kilcoyne A, Wang Y, *et al*. Comparison of the (18)F-FDG avidity at PET of benign and malignant pure ground-glass opacities: a paradox? *Clin Radiol* 2019;**74**:187–95.
12. Boktor RR, Walker G, Stacey R, *et al*. Reference range for intrapatient variability in blood-pool and liver SUV for ^{18}F-FDG PET. *J Nucl Med* 2013;**54**:677–82.
13. Shiono S, Abiko M, Okazaki T, *et al*. Positron emission tomography for predicting recurrence in stage I lung adenocarcinoma: standardized uptake value corrected by mean liver standardized uptake value. *Eur J Cardiothorac Surg* 2011;**40**:1165–9.
14. Tournoy KG, Maddens S, Gosselin R, *et al*. Integrated FDG-PET/CT does not make invasive staging of the intrathoracic lymph nodes in non-small cell lung cancer redundant: a prospective study. *Thorax* 2007;**62**:696–701.
15. MacMahon H, Naidich DP, Goo JM, *et al*. Guidelines for management of incidental pulmonary nodules detected on CT images: from the Fleischner Society 2017. *Radiology* 2017;**284**:228–43.
16. Lawrence J. *Introduction to neural networks*. 5th edn. Nevada City, CA: California Scientific; 1993.
17. Ayer T, Chhatwal J, Alagoz O, *et al*. Informatics in radiology: comparison of logistic regression and artificial neural network models in breast cancer risk estimation. *RadioGraphics* 2010;**30**:13–22.
18. de Bruijne M. Machine learning approaches in medical image analysis: from detection to diagnosis. *Med Image Anal* 2016;**33**:94–7.