Artificial intelligence and radiomics enhance the positive predictive value of digital chest tomosynthesis for lung cancer detection within SOS clinical trial

Artificial intelligence and radiomics enhance the positive predictive value of digital chest tomosynthesis for lung cancer detection within SOS clinical trial

Adriano De Maggi[1], Ilaria Baralis[2], Federico Dalmasso[1], Paola Berchialla[3], Federico Mazza[4], Giulio Melloni[4], Maurizio Grosso[4], Stephane Chauvie[1]

[1]Medical Physics Division, Santa Croce e Carle Hospital, Cuneo, Italy; [2]Epidemiology Depatment, Universitty of Torino, Torino, Italy; [3]Radiology Department, Santa Croce e Carle Hospital, Cuneo, Italy, [3]Thoracic Surgery Department, Santa Croce e Carle Hospital, Cuneo, Italy

Corresponding author: Stephane Chauvie, Santa Croce e Carle Hospital, via Coppino 26 12100 Italy, tel. +39.0171.64.1558, fax +39.0171.64.1554 e-mail chauvie.s@ospedale.cuneo.it

Keyword: digital tomosynthesis, lung cancer, lung nodule detection, random forest

SOS Study team: Alberto Biggi (SC Medicina Nucleare), Andrea Campione, Mirella Fortunato (SC Anatomia Patologica), Adriano De Maggi, Stéphane Chauvie (SC Fisica Sanitaria), Ida Colantonio (SC Oncologia), Maurizio Grosso (SC Radiologia), Giulio Meloni, Federico Mazza, Alessia Stanzi (SC Chirurgia Toracica), Paolo Noceti (SC Pneumologia), Paolo Pellegrino (Direzione Sanitaria), Elvio Russi (SC Radioterapia).

Abstract

Objectives: within this investigation we investigated several approaches to enhance the positive predictive value (PPV) of chest digital tomosynthesis (DTS) in the lung cancer detection

Methods: the investigation was carried out within the SOS clinical trial (NCT03645018) for lung cancer screening with DTS. Lung nodules were identified by visual analysis and then classified using

the diameter and the radiological aspect of the nodule following a modified lung-RADS classification. Haralick texture features were extracted from the segmented nodules. Both semantic variables and radiomics features were used to build a predictive model using two approaches: logistic regression model on a sub-set of variables selected with backward feature selection or machine learning using the whole sub-set of variables. We used two machine learning methods: a Random Forest and a neural network. Machine learning methods were applied to a training set and validated on a test set. Methods were compared using diagnostic accuracy metrics.

Results: binary visual analysis had a good sensitivity (0.95) but a low PPV (0.14). Lung-RADS classification increased the PPV (0.19) but with an unacceptable low sensitivity (0.65). Analogously, logistic regression showed a mildly increased PPV (0.22) and a low sensitivity (0.67). Random Forest demonstrated a low accuracy with a moderate PPV (0.40) but with a dramatically low sensitivity (0.30). Neural network demonstrated to be the best predictor with a nearly perfect PPV (0.95) and a high sensitivity (0.90).

Conclusions: among the various technique to reduce the false positive rates of DTS the neural network demonstrated a very high PPV. The use of visual analysis along with neural network could help radiologists to depict a follow-up strategy after a positive DTS.

Introduction

Lung cancer is the leading cause of cancer-related death around the world. In 1990–2015, there were nearly 2 million new cases per year worldwide accounting for the 11.5% of all new cancer diagnoses and 19.7% of the deaths [1]. Despite decreasing smoking trends in developed countries and resulting decrease in lung cancer mortality, the population at risk for lung cancer continues to be large [2]. Lung cancer screening with low-dose computed tomography (CT) demonstrated, in the pivotal National Lung Screening Trial (NLST) study, a clear reduction in mortality [3]. Analogous results were achieved in succeeding American and European's studies [4]. Our group, starting in 2010, proposed a different approach based on chest digital tomosynthesis (DTS), a radiographic technique that consist of a basculation of the X-rays tube providing three-dimensional images oriented in the coronal plane. On a population of around 2000 subjects we demonstrated a detection rate of lung cancer of 1.1% [5] that is comparable to those obtained in CT screening trials. One of the pitfalls of DTS, shared with CT, is the high rate of false positives. To reduce them we initially proposed a classification of the nodules based on their characteristics [6], and within this work, we compare it against a multi-variate prediction model and a random forest model. The random forest was developed within this work on a independent training population.

Patients and methods

"SOS: Studio OSservazionale" clinical trial (NCT number NCT03645018) enrollment was open between Dec 2010 and Aug 2018, follow-up closed on August 2019. Written informed consent was obtained before entry into the study, in accordance with the requirements of the institutional review board and local health authorities. Subjects considered eligible were smokers or former smokers aged 45 to 75 years, with a smoking history of at least 20 pack-years; for former smokers, the maximum time since smoking cessation was 10 years [5]. DTS were performed at baseline in 2011 (SOS), 1 year after in 2012 (SOS1) and 5 years after in 2017 (SOS2). For every round of DTS, subjects with a non-calcific nodule larger than 5 mm or with multiple nodules were addressed to CT and hence managed in keeping with the most up-to-date Fleischner Society guidelines [7]. Lung cancers were confirmed by histology. Non-lung cancers were confirmed either by histology or by one-year radiological follow-up.

All nodules were classified by at least two independent radiologists following lung-RADS classification [8] based on nodule characteristics: maximum diameter (below 5 mm, between 5 and 8 mm, and higher than 10 mm) and type (solid, sub-solid, Ground-Glass Opacity, GGO). A radiologist and a medical physicist manually contoured together all the nodules using PET Encore (MIM

software Inc., Cleveland, OH, USA) workstation. To reduce the ripple effect on DTS radiomics features were calculated only on the three coronal central slices of the nodule using the open-source and validated PORTS radiomics toolkit [9]. Original pixel's size of 0.2x0.2x3 mm was down-sampled to an isotropic voxel of 3x3x3 mm. Image pixel values were discretized by a fixed number of 64 bins between maximum and minimum to avoid dependence of radiomics texture from pixel values itself. 42 Haralick features [10] were extracted from the images including histogram features and four textural matrices, namely the gray-level co-occurrence matrix (GLCM), neighboring gray-level dependence matrix (NGLDM), gray-level run-length matrix (GLRLM), and gray-level size zone matrix (GLSZM).

Two different approaches were used to evaluate the combined impact of radiomics and nodule characteristics: machine learning (using a Random Forest (RF) and a Neural Network (NNET)) developed on a test set within this work and a logistic regression (LR) method. The RF was trained on SOS2 dataset with pre-identified lung cancers. All radiomics and semantic variables were used in the model to account for the maximum granularity. Each RF was cross validated on 10-fold for 100 times. The goodness-of-fit for RF was estimated with accuracy metrics. Since the SOS2 dataset has few events, i.e. lung cancers, we applied a bootstrap-based oversampling technique [11] to obtain a ratio of events to non-events of 1/3. The model was then applied to SOS dataset.

LR was applied directly to SOS dataset. All radiomics and semantic variables were first analyzed to exclude futility (variables for whom the ratio among the first and the second occurrence was more than 10 were rejected), correlation (variables with a correlation coefficient calculated with Kendall's τ or Spearman's ρ [12] higher than 0.8 were rejected) and linear combination (variables with a existing linear combination among them found with QR decomposition were rejected). Then a backward feature selection method was applied that iteratively found the most important variables among different tested subsets. The remaining variables were fitted by a generalized linear model. The model was then used to predict the lung cancer occurrence in the test dataset (SOS1).
Diagnostics metrics were used to compare the 5 models: 1) the binary scale (positive vs negative) based on visual analysis, 2) the 5-point scale (I to IVb) lung-RADS classification 3) the LR 4) the RF and 5) NNET.
This article was developed following TRIPOD recommendation [13].

Results

1594 subjects were enrolled in the study. 132 and 102 patients were DTS positive at SOS and SOS2, with, respectively 208 and 134 nodules detected by DTS. The characteristic of all the nodules is shown in Table 1. Applying the lung-RADS classification 31, 78, 121, 69 and 43 nodules were in class I, II, III, IVa and IVb, respectively. 20 and 12 lung cancers were found after at least 1-year follow-up at SOS and SOS2.

During feature selection of radiomics and nodule characteristics, 3 variables were rejected for futility, 27 for correlation and none for linear combination. The backward feature selection procedure subsequently applied retained the four variables glzsm_low_gray_level_zone_emphasis, nodule diameter, gtsdm_entropy and glzsm_zone_size_non_uniformity, resulting in a final model that achieved an accuracy higher than 0.85. The final LR model was selected according to the Akaike Information Criteria (AIC = 81) and achieved the Somer's concordance indx Dxy of 0.63 (the closer to 1, the better) and the Brier score of 0.10 (the closer to 0, the better).

SOS2 dataset was used as training. Random Forest had an accuracy of 0.98, a sensitivity of 0.99 and a specificity of 0.98

The most important variables depicted by the RF algorithm were ngtdm_coarseness, nodule diameter, ngtdm_busyness, ngtdm_complexity and ngtdm_texture_strength with a relative importance of 100, 61, 22, 16 and 15, respectively.

SOS dataset was used to compare the diagnostic accuracy of the five methods. 208 nodules were defined as positive by binary visual analysis, 70 (IVa and IVb) with lung-RADS, 18 with LR 15 with RF and 19 with NNET. Diagnostic accuracy of the five models on the test SOS dataset is shown in Table 2.

Discussion

One of the hurdles enduring in lung cancer screening is keeping the positive predictive value as high as possible without compromising the sensitivity. Indeed, false positive findings create discomfort in the patient, loss of confidence in the screening program besides a numerous set of useless examinations. With the binary visual analysis, we indeed found a sensitivity of 0.95 and a specificity of 0.93. A metanalysis [14] of numerous DTS screening trials showed pooled sensitivity of 0.83 and specificity of 0.91, that is in line with our findings. In CT screening adding the lung-RADS classification of the nodules, generally increased the sensitivity but to a cost of higher number of false positive. Pinsky et al. [15] demonstrated an increase of sensitivity from 0.78-0.79 to 0.93-0.94 with a false positive rate raising from 0.05 to 0.16-0.27. In our work, considering positive a nodule with

lung-RADS classes IVa and IVb, we found a sensitivity of 0.65 and a PPV rate of 0.04. While PPV slightly increased we had a great reduction in sensitivity. We shall point-out that at the time of SOS trial start in 2010 we choose the threshold of 5 mm used in Fleischer guidelines and not the 6 mm one used in lung-RADS.

Over the last two decades a plethora of Computer Aided Detection (CAD) systems have been developed to improve diagnostic accuracy in CT. CAD are nowadays generally used as a "second opinion" tool, providing a list of possible lung nodules that shall be characterised by the radiologist. In such setting CAD systems have a high sensitivity (up to 100%) at the cost of a low specificity (up to 8.2 false positive nodules per scan) [16]. At our knowledge there are few experiences on DTS. Dobbins III et al. [17] were the first to develop an automated lung segmentation method and nodule detection. Over a series of 45 DTS they achieved relatively high accuracy for lung segmentation and all of the nodules were correctly found. Hadházi et al.[18] proposed a domain-specific filters for the enhancement and classification of bright, rounded structures along with a vessel enhancing algorithm based on strain energy filters. To reduce false positive findings supervised vector machine-based classifiers were applied, where features obtained from the vessel enhancement module were used as inputs. The system was evaluated on the scans obtained with their experimental DTS system [19]. Of the ~2000 nodule candidates, 97% of them were detected, producing on average 31 false positives per scan.

In this work, instead of developing a CAD system we decided to find a tool to reduce the false positive rate endeavoring the potential role of radiomics. Several experience have been carried out for CT. Balagurunathan et al. [20] used images and data from the NLST, curated a subset of 479 participants (244 for training and 235 for testing) that included lung cancers and nodule-positive controls. After removing redundant and non-reproducible features, optimal linear classifiers with AUC-ROC were used with an exhaustive search approach to find a discriminant set of image features, which were validated in an independent test dataset. They identified several strong predictive models, using size and shape features; the highest AUC was 0.80. Using non-size-based features, the highest AUC was 0.85. Combining features from all the categories, the highest AUC was 0.83. One-hundred-fifty lung nodules among 114 lung cancer patients from the NLST were investigated by Lu et al. [21]. Lung nodules were semi-automatically segmented using lung and mediastinal windows separately, and subtracting the mediastinal window region from the lung window region generated the difference region. The tumor growth could be predicted by radiomic models constructed using features obtained in the lung window, the difference region, and by combining features obtained in both the lung window and difference regions with AUC of 0.80, 0.82, and 0.85, respectively. Wu et al. [22] analyzed radiomics feature for lung cancer detection in a series of 121 subjects. The AUC (and 95%

confidence interval) for the set of radiomics features, for the set of clinical variable and radiological semantics and for the combination of the two sets were 0.85 (0.71–0.96), 0.88 (0.77–0.96), and 0.88 (0.77–0.97), respectively.

At our knowledge the radiomics of lung nodules was never studied DTS. Within this work we used two different approach. One using linear regression model to fit the lung cancer based on the variables identified though analytical and backward feature selection and the other one with predictive model based on two machine learning technique: a random forest and a neural network. The machine learning algorithms were trained on SOS2 dataset and tested on SOS. Both radiomics features and nodule characteristics (nodule diameter and type) were used. The features selected by the algorithm were the nodule dimension, since, even if obvious, big nodules are more likely to be lung cancers, gtsdm_entropy that is a measure of nodule dis-organization and glzsm_low_gray_level_zone_emphasis and glzsm_zone_size_non_uniformity that account for areas with different grey intensity within the nodules. The LR permitted to increase the PPV respect to visual analysis to 0.22 but with a still too low sensitivity 0.67. Similar effect we found when we used the RF algorithm. Even if the PPV raised to 0.40 the price to pay was the sensitivity dropping down to 0.30. On the other hand the predictive value of neural network proved to be impressive. The PPV jumped to 0.95 but with a sensitivity comparable to that of visual analysis (0.90).

We know that an intrinsic limitation of DTS, due to the limited angle reconstruction, is the ripple artifacts in the antero-posterior direction. Consequently, the radiomics features could be diluted and the intrinsic heterogeneities of the tumor, disguised. That's the reason why we calculated radiomics indexes only on the three central slices in the coronal plane. As a future work we are considering adding some shape radiomic features that could, as already seen with CT add some additional information. One of the major limitations of the study is the few numbers of events. To overcome it we could have used an independent test set of all the lung cancers discovered in DTS in our hospital, outside the SOS clinical trial. Although, they were few because CT, and not DTS, is used as first-line diagnostic tool in subjects with a suspect of lung cancer.

Conclusions

In this work we tried to introduce different techniques to increase the positive predictive value of digital chest tomosynthesis in lung cancer detection. Considering the different radiological appearance of nodule in CT and DTS the lung-RADS classification did not add diagnostic accuracy

to visual analysis. Among the other techniques, neural network was the only one to have a great PPV with loosing sensitivity.

Acknowledgments: A special thanks to the technologists Kawtar Nourani and Denise Guerra for their precious collaboration.

Funding: Grant support for the SOS clinical trial was provided by "Cassa di Risparmio di Cuneo" Foundation. Santa Croce e Carle, the hospital where the study was performed provided logistic support, telephone lines, software, computer assistance, and an office free of charge.

Conflict of interests: the authors declare that they have not conflict of interests.

Ethical approval: all applicable international, national and institutional guidelines for the care and use of animals were followed. All procedure performed in studies involving human participants were in accordance with the ethical standards of the institution and national research committee and with the 1964 Helsinki declaration and its later amendments of comparable ethical standards. Study was approved by local IRB. Informed consent was obtained from all individual participants included in the study.

Bibliography

1. Fitzmaurice C, Allen C, Barber RM, et al (2017) Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 32 cancer groups, 1990 to 2015: A Systematic Analysis for the Global Burden of Disease Study Global Burden . JAMA Oncol 3:524–548. https://doi.org/10.1001/jamaoncol.2016.5688
2. Aberle DR, Adams AM, Berg CD, et al (2011) Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 365:395–409
3. Church TR, Black WC, Aberle DR, et al (2013) Results of initial low-dose computed tomographic screening for lung cancer. N Engl J Med 368:1980–91. https://doi.org/10.1056/NEJMoa1209120
4. Bach PB, Mirkin JN, Oliver TK, et al (2012) Benefits and harms of CT screening for lung cancer: A systematic review. JAMA - J Am Med Assoc 307:2418–2429. https://doi.org/10.1001/jama.2012.5521
5. Terzi A, Bertolaccini L, Viti A, et al (2013) Lung cancer detection with digital chest tomosynthesis: baseline results from the observational study SOS. J Thorac Oncol 8:685–92. https://doi.org/10.1097/JTO.0b013e318292bdef
6. Grosso M, Priotto R, Ghirardo D, et al (2017) Comparison of digital tomosynthesis and computed tomography for lung nodule detection in SOS screening program. Radiol Medica 122:568–574. https://doi.org/10.1007/s11547-017-0765-3
7. MacMahon H, Austin JHM, Gamsu G, et al (2005) Guidelines for management of small pulmonary nodules detected on CT scans: a statement from the Fleischner Society. Radiology 237:395–400
8. McKee BJ, McKee AB, French R, et al (2012) “Lung-rads” a proposed standardized reporting and data system for ct lung cancer screening. J Thorac Oncol 4):S277–S278
9. Hatt M, Tixier F, Pierce L, et al (2017) Characterization of PET / CT images using texture analysis : the past , the present … any future ? Eur J Nucl Med Mol Imaging 151–165. https://doi.org/10.1007/s00259-016-3427-0
10. Haralick RM, Dinstein I, Shanmugam K (1973) Textural Features for Image Classification. IEEE Trans Syst Man Cybern. https://doi.org/10.1109/TSMC.1973.4309314
11. Menardi G, Torelli N (2014) Training and assessing classification rules with imbalanced data. Data Min Knowl Discov. https://doi.org/10.1007/s10618-012-0295-5
12. KENDALL MG (1945) THE TREATMENT OF TIES IN RANKING PROBLEMS. Biometrika 33:239–251. https://doi.org/10.1093/biomet/33.3.239
13. Moons KGM, Altman DG, Reitsma JB, et al (2015) Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. Ann Intern Med 162:W1–W73. https://doi.org/10.7326/M14-0698
14. Kim JH, Lee KH, Kim KT, et al (2016) Comparison of digital tomosynthesis and chest radiography for the detection of pulmonary nodules: Systematic review and meta-analysis. Br J Radiol 89:. https://doi.org/10.1259/bjr.20160421
15. Pinsky PF, Gierada DS, Black W, et al (2016) Performance of Lung-RADS in the National Lung Screening Trial. https://doi.org/10.7326/M14-2086
16. Al Mohammad B, Brennan PC, Mello-Thoms C (2017) A review of lung cancer screening and the role of computer-aided detection. Clin Radiol 72:433–442. https://doi.org/10.1016/j.crad.2017.01.002
17. Wang J, Dobbins JT, Li Q (2012) Automated lung segmentation in digital chest tomosynthesis. Med Phys 39:732–741. https://doi.org/10.1118/1.3671939
18. Hadhazi D, Varga R, Horvath A, et al (2015) Digital chest tomosynthesis: The main steps to a computer assisted lung diagnostic system. In: 2015 IEEE International Symposium on Medical Measurements and Applications, MeMeA 2015 - Proceedings. pp 40–45

19. Horváth Á, Wolf P, Nagy J, et al (2016) OVERVIEW OF A DIGITAL TOMOSYNTHESIS DEVELOPMENT: NEW APPROACHES FOR LOW-DOSE CHEST IMAGING. Radiat Prot Dosimetry 169:171–176. https://doi.org/10.1093/rpd/ncv469
20. Balagurunathan Y, Schabath MB, Wang H, et al (2019) Quantitative Imaging features Improve Discrimination of Malignancy in Pulmonary nodules. Sci Rep 9:1–14. https://doi.org/10.1038/s41598-019-44562-z
21. Lu H, Mu W, Balagurunathan Y, et al (2019) Multi-window CT based Radiomic signatures in differentiating indolent versus aggressive lung cancers in the National Lung Screening Trial: A retrospective study. Cancer Imaging 19:. https://doi.org/10.1186/s40644-019-0232-6
22. Wu W, Pierce LA, Zhang Y, et al (2019) Comparison of prediction models with radiological semantic features and radiomics in lung cancer diagnosis of the pulmonary nodules: a case-control study. Eur Radiol. https://doi.org/10.1007/s00330-019-06213-9

	Solid	*Partially solid*	*Ground Glass Opacity*	*Calcific*	*All*
<5 mm	49	0	1	12	*62*
5-8 mm	120	11	1	11	*143*
8-10 mm	47	2	2	3	*54*
>10 mm	62	16	3	2	*83*
All	*278*	*29*	*7*	*28*	*342*

Table 1 Distribution of nodules' diameters and type in SOS1 and SOS3 combined data set

	Sensitivity	*Specificity*	*Positive Predictive Value*	*Negative Predictive Value*	*True Positive*	*True Negative*	*False Positive*	*False Negative*
Binary Visual Analysis	0.95 (0.75; 1.00)	0.93 (0.91; 0.94)	0.14 (0.09; 0.21)	1.00 (1.00;1.00)	19	1460	114	1
lung-RADS classification	0.65 (0.41; 0.85)	0.96 (0.95; 0.97)	0.19 (0.10; 0.30)	1.00 (0.99; 1.00)	13	1517	57	7
Logistic Regression	0.67 (0.22; 0.96)	0.99 (0.99; 1.00)	0.22 (0.06; 0.48)	1.00 (1.00;1.00)	4	1574	14	2
Random Forest	0.30 (0.12; 0.54)	0.99 (0.99; 1.00)	0.40 (0.16; 0.68)	0.99 (0.99; 1.00)	6	1565	9	14
Neural network	0.90 (0.68; 0.99)	1.00 (1.00;1.00)	0.95 (0.74; 1.00)	1.00 (1.00; 1.00)	18	1573	1	2

Table 2 Diagnostic accuracy metrics for the different classification algorithms: p*oint estimates and 95%* confidence interval