Deep learning analysis of lung nodules decreases false positive rate in the National Lung Screening Trial

# Introduction

Low-dose Computed Tomography (LDCT) images collected during epidemiological studies, clinical trials, and routine screening represent an opportunity to improve the accuracy of cancer risk predictions. LDCT imaging is an effective modality for early detection and diagnosis of lung cancer. The National Lung Screening Trial (NLST) reported a 20% reduction in mortality in the LDCT arm relative to the standard X-ray arm.1

The use of LDCT for lung cancer screening is not without limitations, however, and current obstacles include - high inter-rater variability, i.e. differences among the assessments of different clinicians and - high false positive rates, i.e. detecting cancer in individuals without cancer. In the NLST, for example, roughly a quarter (26.6%) of the positive baseline LDCT scans did not result in a cancer diagnosis.1 Methods to improve the false positive rate have been successfully implemented,2 but false positives remain a major concern.

In an effort to overcome the limitations described above, we aim to develop lung cancer risk prediction models that can integrate LDCT image features with established risk factors. Current statistical models used to assess cancer risk focus on well-described risk factors, such as smoking, and demographic factors, like race.3 Integrating image data into statistical models represents a significant challenge, because of the computational requirements and technical expertise required to extract image features as numeric variables that can be used in statistical models.

As part of this work, we will develop an image analysis pipeline that trains three-dimensional convolutional neural networks (3D CNN) to distill LDCT lung scans into variables for lung cancer risk prediction models. A 3D CNN utilizes multiple layers of convolutions to detect features all throughout a lung scan, as opposed to a two-dimensional CNN which extracts feature from a single image slice at a time. 3D CNN have the potential to yield improved results, but require far greater computational resources than 2D CNN.

Algorithmic feature extraction from LDCT images offers many advantages over manual annotation of images by physicians. In particular, CNN image feature extraction is quantitative and can be automated to as part of systems that supply image features for downstream lung cancer risk prediction models or that output predicted lung cancer risk directly. These images features can also be used to identify conditions that may be associated with lung cancer risk, like emphysema.

The work described here will constitute a proof of principle of cancer risk modeling methods that combine algorithmic imaging analysis techniques with classical cancer risk modeling. Algorithmic feature extraction from LDCT images offers many advantages over manual annotation of images by physicians. In particular, algorithms can be automated and yield quantitative results, whereas image annotation by physicians is tedious and difficult to quantify. Furthermore, the reproducibility and reliability of algorithmic approaches can be ensured through software engineering best practices.

While our current focus is lung cancer, the methods we will develop could potentially be adapted for other cancer types. For example, it may be possible employ a similar approach using image data from other medical screening modalities, such as mammograms to improve breast cancer risk prediction or colposcopic images to assess cervical cancer risk. Similarly, we can extend the utility of our proposed approach beyond cancer by shifting our focus to LDCT image characteristics related to heart disease and chronic obstructive pulmonary disease. The ultimate goal of this work is to highlight the importance of screening, inform screening guidelines, and provide individuals with improved assessments of their health risks.

Integrating heterogeneous data into statistical models represents a significant technical challenge. The work described here constitutes a proof of principle of a cancer risk modeling method that combines imaging analysis techniques with classical cancer risk modeling. Current statistical models used to assess cancer risk focus on well-described risk factors, such as smoking, and demographic factors, such as race and ethnicity.

Low-dose Computed Tomography (LDCT) images collected during epidemiological studies, clinical trials, and routine screening represent an opportunity to improve the accuracy of cancer risk predictions. LDCT imaging is an effective modality for early detection and diagnosis of lung cancer. The use of LDCT for lung cancer screening is not without limitations, however, and current obstacles include 1) high inter-rater variability, i.e. differences among the assessments of different clinicians and 2) high false positive rates, i.e. detecting cancer in individuals without cancer.

In an effort to overcome these limitations, we aim to develop lung cancer risk prediction models that can integrate physician-annotated and algorithmically derived LDCT image features with established risk factors, such as smoking exposure. In essence, our approach is to update prescreening risk, calculated based on known risk factors, with information obtained from LDCT screening. To evaluate the feasibility and utility of this approach, we analyzed LDCT images obtained from the National Lung Screening Trial (NLST).

We used an AI developed for discrimination of malignant and benign nodules, to see whether it could improve a clinical model for the prediction of a screening patient’s likelihood of developing cancer over the upcoming year. The existing model used a number of features extracted from the metadata of the US National Lung Screening Trial (NLST) to produce a likelihood score, given a screening output in year *y*, of developing cancer before the corresponding screening image taken at year *(y+1)*. These features included both patient clinical data (age, smoking history etc), and quantities extracted from a radiological read of the screening CT, such as maximal nodule size, the existence of any GGO nodules on the CT, and presence of emphysema within the patient.

# Methods

To evaluate the feasibility and utility of our proposed approach and follow through with aims described in this proposal, we analyzed LDCT images obtained from the National Lung Screening Trial (NLST). We obtained LDCT scans from all of the participants (n = 26722) in the LDCT arm of the NLST.1 Each study participant was scheduled to receive three yearly scans. LDCT scan were labeled by radiologists at the site of the screening for lung abnormalities. We developed data processing pipelines in the R programming language to integrate clinical and epidemiologic data from the NLST with features that we extracted from the LDCT scan scan images using 3D CNNs.

Before the LDCT scans can be used to train 3D CNNs, the LDCT scans must processed. To generate processed data, the LCDT scans were first converted to NifTI-1 format, then cropped to a bounding box around the lung using the Progressive Holistically-Nested Network (P-HNN) lung segmenter4, normalized in the three different lung windows of [−1000, 200], [−160, 240], and [−1000, −775] Hounsfield units, and rescaled to a standard size of 128x128x128. The resulting 3-channel images then be used for 3D CNN training.

Our proposed neural network architecture consists of five 4-layer blocks of 3x3x3 convolution, batch normalization, ReLU activation, and 3D max pooling; then a convolution group of 2x2x2 convolution, batch normalization, Rectified Linear Unit (ReLU) activation, and 50% dropout before a fully connected group of 1x1x1 convolution, 50% dropout, 1x1x1 convolution, 50% dropout, a flattening layer, and a dense layer with 2 class outputs. The labels used for training are physician annotations of lung conditions. The loss function we used for binary classification of lung health conditions, for example emphysematous or non-emphysematous, is cross entropy. All of our 3D CNNs are implemented with the Python programming language and the Tensorflow5 library.

Previously, our group has shown that prescreening risk, calculated with the Lung Cancer Risk Assessment Tool (LCRAT) model,3 can be combined with physician annotations of lung conditions to more accurately predict lung cancer risk.6 In this previous work, we used log-binomial regression7 to fit first-order Markov transition models.8 Rather than using physician annotations in a model directly, we used physician annotations of lung conditions to train 3D CNNs to produce quantitative estimates of lung condition severity. We then used a statistical technique called stacking to integrate LCRAT-calculated prescreening risk with 3D CNN outputs. Stacking aggregates the multiple models outputs using a combiner algorithm that can described as a meta-learner.9–11 In summary, we propose a novel lung cancer risk prediction method that utilizes a meta-learning algorithm to combine the outputs of the LCRAT model and 3D CNNs trained to identify lung conditions.

The chest CT scans used in this study were taken from the all three years of the NLST from both the LSS branch and the ACRIN branch; 8416 patients were used for training, 480 for validation, and 2016 for testing. Each CT scan was labeled by radiologists at the site of the screening for lung CT abnormalities.

A re-curated version of the NLST dataset CTs and metadata was used for this study. Each CT listed as containing at least one nodule was reviewed by a medical doctor or medical student, under expert supervision from University of Oxford Radiologists, and metadata records of the CT findings were reviewed and extended. In particular, the size, extent, location, margins and attenuation of each nodule were reviewed by the same small team of individuals, and the exact 3D location of each nodule was identified and recorded. Additional nodules not listed in the

NLST metadata were also added as long as they were not fully calcified (since fully calcified nodules were not considered “positive findings” in the original NLST data). As well as reviewing all CTs on which one or more nodules was recorded, the team also reviewed all CTs of patients recorded as having developed lung cancer, and again fully reviewed and extended their mark-up and metadata. Patients who never had any reported nodules and also never had lung cancer were not considered or marked up.

An AI called the LCP-CNN (Lung Cancer Prediction Convolutional Neural Network) was trained on this augmented NLST set in an 8-fold cross-validated way using Tensorflow.5 The AI training is outside the scope of this publication, but for reference, the training used class balancing, and also extensive pre-training on hundreds of thousands of non-NLST images in order to achieve strong performance both on the testing folds of the cross-validation, and also on several independent external datasets.

The AI was trained for the task of malignant vs benign classification, and produces a score where 0 indicates benignity, and 100 indicates malignancy.

The set of features extracted from the NLST metadata and the re-curated CT information represents the full space of information available to our new model.

Table of each feature under consideration for the screening models. Detail which features are derived from where. LP or ND will use this list to produce the up-to-date release of Optellum’s nodule data and metadata for use in this project.

The new model we are fitting is an extension of <background description of Hilary’s work>. The particular form of the model being fitted is <equations>.

Two versions of the model were fitted using the feature set described above. In the first, the maximum LCP score for a given CT was *not* available as a feature, and in the second, it was available for selection. In both cases, feature selection was done using Bayes Information Criterion or Lasso.

The four models were then compared according to TBD the necessary but sufficient statistical tests, showing that in both cases the addition of the LCP-CNN gave a better fit to data.

# Results

We should defer writing up any results until we’ve reached consensus with everything above. most of this is required during training and not during deployment.

# Discussion

The results presented here show that 3D CNN-based approaches are a promising direction of research for computer-aided prediction of a future cancer diagnosis. As well as just diagnosing cancer, neural networks could identify multiple CT-diagnosed diseases to identify risk factors for lung cancer, or even identify lung cancer risk directly. Furthermore, the neural network may be useful as significantly more robust than the density mask approach, even with a very basic network.

When applied to data obtained from Low-Dose Computed Tomography (LDCT) lung cancer screening, a Convolutional Neural Network (CNN) can yield information on entire organs, like the lungs or heart, or specific regions of interest, such as any existing lung cancer nodules. Essentially, CNNs are algorithms that distill complex data, typically images, into simpler outputs. Each CNN must be trained to perform a specific task. In particular, we trained CNNs to generate quantitative estimates of the severity of - heart conditions like coronary artery calcification, and lung conditions like emphysema, adenopathy, and consolidation. We hypothesize that these health condition estimates will allow us to gain valuable insight into the risks of lung cancer, chronic obstructive pulmonary disease, and heart disease.

In essence, our approach is to update prescreening risk, calculated using standard lung cancer risk prediction models, with information derived from LDCT screening by CNNs. We have previously demonstrated the utility of a similar approach that used physician annotations of LDCT screens. For example, we have shown that LDCT scans annotated by physicians as emphysematous are associated with increased lung cancer risk. Building on our previous work, we endeavor to develop a model that will aggregate the results of a prescreening risk model and multiple CNNs into a single lung cancer risk prediction. We hypothesize that a model which uses image features provided by a CNN instead of physician-annotated features will similarly result in a better risk assessment than prescreening risk alone, but without the need for physicians to manually annotate LDCT scans.

Neural networks are often referred to as “black box” models, because it is very difficult to explain exactly what inputs result in a given output. We can use a technique called local interpretable model-agnostic explanations to pinpoint exactly what parts an image influence cancer risk, but this information is still be difficult to interpret. In contrast, if we obtain an estimate of lung condition severity from image data, we can then use interpretable statistical models to analyze how that lung condition affects lung cancer risk. In essence, the approach we are proposing combines the interpretability of classical statistical modeling with the ability of neural networks to extract information from images. We hypothesize that this combined model will have comparable predictive performance to a CNN that predicts lung cancer risk directly, but will additionally help to elucidate the factors that influence predictions.

In the case of LDCT scans with identifiable lung cancer nodules, CNNs can provide a quantitative malignancy score based on nodule features. Malignancy scores can help distinguish between malignant and benign nodules, inform prognosis, and avoid unneeded interventions in response to the discovery of insignificant nodules. This approach focuses on the nodules and thus loses the context of the surrounding lung. We propose an alternate approach that uses a malignancy score produced by a CNN in addition to prescreening risk and information on lung condition obtained from physician annotation or using a CNN. We hypothesize that by taking into account prescreening risk and the state of the surrounding lungs we can avoid more false positives than with a CNN-derived malignancy score alone.

# Tables

Table 1: Model Table

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| model | AIC | auroc | gain\_cap | deviance | lrt\_p\_value |
| lcrat\_ct | 1635.449 | 0.7774042 | 0.5548084 | 1613.449 | NA |
| lcp | 1520.130 | 0.8500406 | 0.7000813 | 1516.130 | NA |
| lcp\_lcrat | 1513.982 | 0.8506717 | 0.7013435 | 1507.982 | 0.0043123 |
| lcp\_lcrat\_ct | 1497.282 | 0.8596483 | 0.7192966 | 1473.282 | 0.0000673 |

Table 2: Variable Table

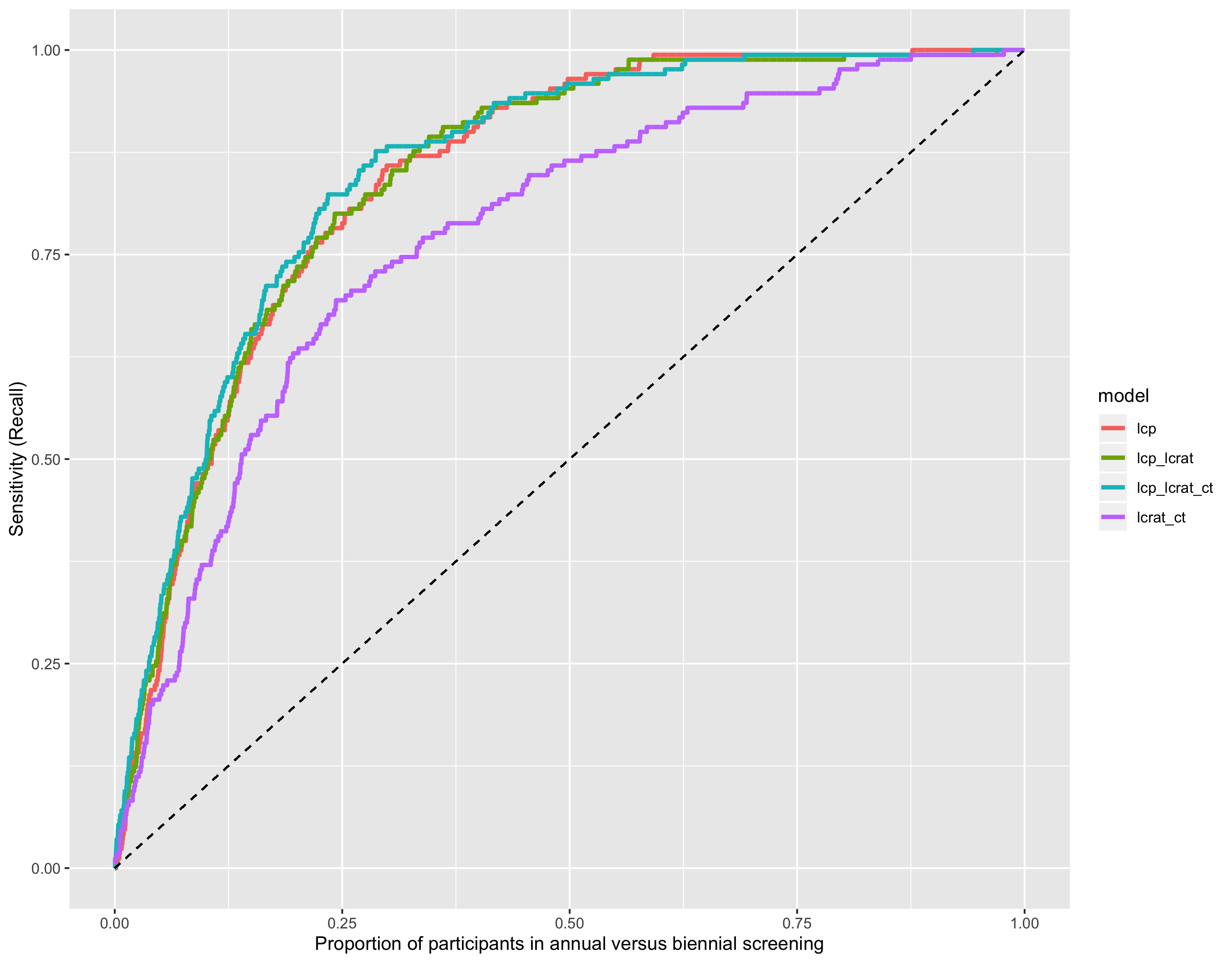
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| model | variable | coefficient | standard\_error | p\_value |
| LCP | (Intercept) | -5.4665494 | 0.1471436 | 0.0000000 |
| LCP | max\_lcp\_score | 3.5033427 | 0.2193836 | 0.0000000 |
| LCRAT+CT | (Intercept) | -3.2629968 | 0.6019510 | 0.0000001 |
| LCRAT+CT | logit1yrisk | 0.3520113 | 0.1046703 | 0.0007709 |
| LCRAT+CT | longest\_diam | 0.0144631 | 0.0064472 | 0.0248769 |
| LCRAT+CT | any\_growth1 | 0.9622552 | 0.1974392 | 0.0000011 |
| LCRAT+CT | any\_upper | 0.1863720 | 0.1709962 | 0.2757490 |
| LCRAT+CT | any\_right\_mid | -0.5006827 | 0.2259791 | 0.0267179 |
| LCRAT+CT | any\_lingula | -0.3148836 | 0.3213983 | 0.3272194 |
| LCRAT+CT | any\_mixed | 0.1123184 | 0.2295399 | 0.6246153 |
| LCRAT+CT | any\_spiculation | 1.1367254 | 0.1812429 | 0.0000000 |
| LCRAT+CT | any\_poor\_def | 0.9594292 | 0.1781344 | 0.0000001 |
| LCRAT+CT | any\_margin\_unab | 1.2218871 | 0.2706616 | 0.0000063 |
| LCP+LCRAT | (Intercept) | -3.8138634 | 0.5893008 | 0.0000000 |
| LCP+LCRAT | max\_lcp\_score | 3.4034547 | 0.2218613 | 0.0000000 |
| LCP+LCRAT | logit1yrisk | 0.2973540 | 0.1042624 | 0.0043448 |
| LCP+LCRAT+CT | (Intercept) | -4.0729571 | 0.6316830 | 0.0000000 |
| LCP+LCRAT+CT | logit1yrisk | 0.2774090 | 0.1069117 | 0.0094660 |
| LCP+LCRAT+CT | longest\_diam | -0.0214797 | 0.0110000 | 0.0508544 |
| LCP+LCRAT+CT | any\_growth1 | 0.4993806 | 0.2024922 | 0.0136566 |
| LCP+LCRAT+CT | any\_upper | 0.0956172 | 0.1722511 | 0.5788239 |
| LCP+LCRAT+CT | any\_right\_mid | -0.3477912 | 0.2299525 | 0.1304200 |
| LCP+LCRAT+CT | any\_lingula | 0.0134458 | 0.3246836 | 0.9669675 |
| LCP+LCRAT+CT | any\_mixed | -0.0801330 | 0.2328975 | 0.7307939 |
| LCP+LCRAT+CT | any\_spiculation | 0.4473534 | 0.1973262 | 0.0233856 |
| LCP+LCRAT+CT | any\_poor\_def | 0.7496142 | 0.1865460 | 0.0000586 |
| LCP+LCRAT+CT | any\_margin\_unab | 0.9465197 | 0.2815364 | 0.0007739 |
| LCP+LCRAT+CT | max\_lcp\_score | 3.1631804 | 0.2742864 | 0.0000000 |

Table 3: Risk Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| threshold | x\_absolute | x\_percent | y\_absolute | y\_percent |
| 0.00421 | 259 | 2.217276 | 0 | 0.000000 |
| 0.00500 | 6323 | 54.130640 | 11 | 6.470588 |
| 0.01000 | 8787 | 75.224724 | 37 | 21.764706 |
| 0.02000 | 9813 | 84.008218 | 59 | 34.705882 |
| 0.04000 | 10494 | 89.838199 | 87 | 51.176471 |
| 0.08000 | 11144 | 95.402791 | 131 | 77.058824 |
| 0.12000 | 11648 | 99.717490 | 169 | 99.411765 |

# Figures

Figure 1: Lorenz curve



1. Reduced lung-cancer mortality with low-dose computed tomographic screening. *New England Journal of Medicine*. 2011;365(5):395-409. doi:[10.1056/NEJMoa1102873](https://doi.org/10.1056/NEJMoa1102873)

2. Pinsky PF, Gierada DS, Black W, et al. Performance of lung-RADS in the national lung screening trial. *Annals of Internal Medicine*. 2015;162(7):485. doi:[10.7326/m14-2086](https://doi.org/10.7326/m14-2086)

3. Katki HA, Kovalchik SA, Petito LC, et al. Implications of nine risk prediction models for selecting ever-smokers for computed tomography lung cancer screening. *Annals of Internal Medicine*. 2018;169(1):10. doi:[10.7326/m17-2701](https://doi.org/10.7326/m17-2701)

4. Harrison AP, Xu Z, George K, Lu L, Summers RM, Mollura DJ. Progressive and multi-path holistically nested neural networks for pathological lung segmentation from ct images. In: *MICCAI 2017*.; 2017:1-8.

5. Chollet F, others. Keras. 2015.

6. Robbins HA, Berg CD, Cheung LC, Chaturvedi AK, Katki HA. Identification of candidates for longer lung cancer screening intervals following a negative low-dose computed tomography result. *JNCI: Journal of the National Cancer Institute*. 2019;111(9):996-999. doi:[10.1093/jnci/djz041](https://doi.org/10.1093/jnci/djz041)

7. Wacholder S. Binomial regression in glim: Estimating risk ratios and risk differences. *American Journal of Epidemiology*. 1986;123(1):174-184. doi:[10.1093/oxfordjournals.aje.a114212](https://doi.org/10.1093/oxfordjournals.aje.a114212)

8. Diggle P, Diggle PJ, Heagerty P, et al. Analysis of longitudinal data. In: 2nd ed. Oxford University Press; 2013:190-206.

9. Wolpert DH. Stacked generalization. *Neural Networks*. 1992;5:241-259.

10. Breiman L. *Machine Learning*. 1996;24(1):49-64. doi:[10.1023/a:1018046112532](https://doi.org/10.1023/a:1018046112532)

11. Laan MJ van der, Polley EC, Hubbard AE. Super learner. *Statistical Applications in Genetics and Molecular Biology*. 2007;6(1). doi:[10.2202/1544-6115.1309](https://doi.org/10.2202/1544-6115.1309)