[原著論文]

Radiomics を用いた非小細胞肺がんの EGFR 遺伝子変異の推定 ー遺伝子発現パターンの違いが表現型に及ぼす影響ー

藏本　裕香[†]・内山　良一[††*]

[†]熊本大学大学院保健学教育部　〒862-0976 熊本県熊本市中央区九品寺 4-24-1
[††]熊本大学大学院生命科学研究部　〒862-0976 熊本県熊本市中央区九品寺 4-24-1
*責任著者：内山　良一
（受理日：2021 年 5 月 25 日，採択日：2021 年 9 月 4 日）

Estimation of EGFR gene mutation in non-small cell lung cancers by using radiomics : Effect of differences in gene expression patterns

Yuka KURAMOTO[†] and Yoshikazu UCHIYAMA[††*]

[†]Graduate School of Health Sciences, Kumamoto University, 4-24-1 Kuhonji, Chuo-ku, Kumamoto, Kumamoto 862-0976, Japan
[††]Department of Medical Image Sciences, Faculty of Life Sciences, Kumamoto University, 4-24-1 Kuhonji, Chuo-ku, Kumamoto, Kumamoto 862-0976, Japan
*Corresponding author : Yoshikazu UCHIYAMA
（Received on May 25, 2021. In final form on September 4, 2021.）

Abstract : Since cell becoming cancerous processes due to the accumulation of gene mutations, the effect of gene mutations other than EGFR gene might be affected imaging phenotypes of lung cancer. The purpose of this study is to clarify problems in estimating EGFR gene mutation in lung cancer by using non-invasive image examination. We collected 119 CT images and 18 RNA-Seq data from the NSCLC-Radiogenomics database and conducted experiments. The region of lung cancer was manually segmented and 365 radiomic features were determined. Linear discriminant analysis with 10 radiomic features selected by Lasso was employed for estimating the presence or absence of EGFR gene mutation. In addition, 18 RNA-Seq data with EGFR gene mutation was projected into two-dimensional space by using t-SNE. Experimental results showed that lung cancers with EGFR gene mutation have two groups with different imaging phenotypes. The patterns of gene expression level in those groups were also different. In the radiomic studies for estimating the presence or absence of EGFR gene mutation, we can conclude that it is appropriate to conduct it as a multi-group classification problem incorporated differences of gene expression pattern in lung cancer.
Keywords : Radiomics, Lung cancer, EGFR, CT image

1. 緒　言

細胞の遺伝子変異は，喫煙，化学物質，放射線，紫外線などの環境因子によって生じる．通常，細胞はそれらの変異を修復する機能を備えているが，加齢などによって修復する力が弱まると，変異が蓄積されていき，がん細胞が発生する．このようながん細胞が増殖を重ねることで，がんの形成が進む[1]．近年，分子生物学的な研究が目覚ましい発展を遂げ，肺がんの増殖や転移に関わる遺伝子変異が明らかになり，それらの遺伝子変異を標的とし，働きを阻止する分子標的薬も開発されている[2]．これまで肺がんは，病理組織像を用いて非小細胞肺がんと小細胞肺がんに大別され，治療方針が決定されてきた．しかし，発がんに関わる遺伝子変異が発見されたことで，がんの遺伝型に基づいた個別化医療が進展し，治療の形が急速に変貌しつつある[1, 2]．

EGFR 遺伝子変異は，肺がんにおいて高頻度で発見される変異のひとつであるため，その変異を有する肺がんに対する EGFR 阻害剤が開発され，分子標的薬による治療が行われている．EGFR は細胞の表面にある受容体であり，EGF と結合することで細胞の成長と増殖の調節の役割を担っている．しかし，遺伝子変異によって調節機能が働かなくなると増殖に歯止めが効かなくなりがん化が進む．細胞増殖が進めば，腫瘍の表現型（画像所見）に影響を及ぼすことは容易に想像できるため，腫瘍の Radiomics 特徴量（大きさ，形状，テクスチャなどの画像特徴量）を用いて EGFR 遺伝子変異の有無を推定する研究が行われている[3-10]．

画像を用いて非侵襲に肺がんの EGFR 遺伝子変異ありが特定されれば，生検で細胞の採集が困難な位置に肺がんが存在する場合でも効果的な分子標的薬 EGFR 阻害剤を選択することができる．しかし，上述したように，遺伝子変異の蓄積によってがん細胞が発生するため，EGFR 遺伝子変異以外の遺伝子変異の影響がある場合には EGFR 阻害剤の効果に影響を及ぼす可能性がある．その場合は，肺がんの表現型から EGFR 遺伝子変異ありの情報を取り出し推定できたとしても，EGFR 阻害剤で奏効しない患者も含めて選別することになるため，Radiomics の臨床的な有用性が半減する．

これまでの先行研究[3-10]において，画像から EGFR 遺伝子変異の有無を推定することが可能であることは示されているが，さらに 1 歩先に進めて EGFR 阻害剤の効果が高い群を選別できるかについて，その可能性を検討した研究は我々の知る限り行われていない．そこで本研究では，肺がんの Radiomics 特徴量を用いて EGFR 遺伝子変異の有

無を推定する手法を構築し，その推定結果と腫瘍の遺伝子発現パターンを比較することによって，肺がんの表現型から EGFR 遺伝子変異の有無を推定する際の問題点を明確にする．

2. 方　法

2.1　実験試料

本研究では，The Cancer Imaging Archive の NSCLC-Radiogenomics データベースを用いた[11]．このデータベースには，非小細胞肺がん患者 211 例のデータが収録されている．このうち，ステージ 0 の症例を除外し，EGFR 変異の有無が確定している 119 名の治療前 CT 画像を選択して実験に用いた．EGFR 遺伝子変異ありが 24 症例，EGFR 遺伝子変異なしが 95 症例である．CT 画像のマトリクスサイズは 512×512，ピクセルサイズは 0.6～1.0 mm，スライス厚は 0.625～3.0 cm であった．また，同データベースには，RNA-Seq の情報も公開されている．本研究では，EGFR 遺伝子変異ありの患者 24 名のうち, RNA-Seq のデータが存在する 18 名を選択して実験を行った．このデータには 22126 個の遺伝子発現量があるが，その多くが NA であった．選択した 18 名の発現量がすべて NA のものは解析から除き，18 名×17302 個の遺伝子発現量からなる表を遺伝子データとして用いた．なお，本研究の実施にあたり，倫理審査委員会の承認を得ている．

2.2　Radiomics 特徴量の計測

データベースから取得した 119 症例のすべての治療前 CT 画像に対して，複数枚あるスライス画像から腫瘍の面積が最大となるスライス画像を 1 枚選択した．次に，腫瘍の領域を手動でマーキングした．この際に，がんの辺縁にあるスピキュラなどの形状特徴が正確に計測できるようにマーキングを行った．Fig.1に腫瘍領域のマーキング結果の例を示す．

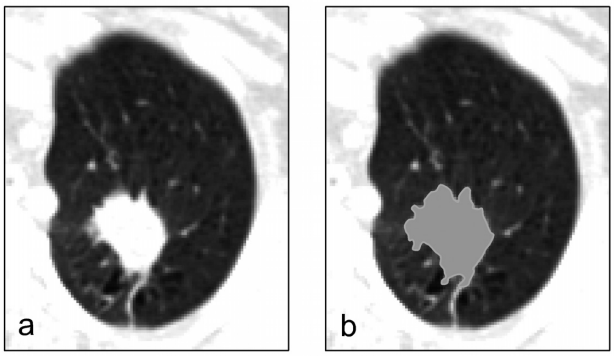

Fig.1 An example of manually segmented tumor region. (a) original image. (b) segmented region.

マーキングした腫瘍領域から 365 項目の Radiomics 特徴量を計測した．Radiomics 特徴量の計測には，一般公開されている MaZda[12-14]を用いた．365 項目の Radomics 特徴量の内訳は，形状 72 個，ヒストグラム特徴量 9 個，テクスチャ 272 個，解像度に関する特徴量が 12 個である．Radiomics 特徴量を計測する際のパラメータは，MaZda のデフォルト値を用いた．例えば，テクスチャ特徴量を計測する際の濃度共起行列を計算する際のパラメータは，濃度階調が 16 ビット，画素間の距離は 1～5，方向は 0 度，45 度，90 度，135 度である．

2.3　Radiomics 特徴量の選択

Radiomics 特徴量が 365 項目であり，実験に用いた症例数が 119 症例であるため，Radiomics 特徴量の次元削減を Lasso（least absolute shrinkage and selection operator）[15]を用いて行った．

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \quad (1)$$

ここで，y_i は i 番目の患者の EGFR 遺伝子変異の有無，x_j は Radiomics 特徴量の値，β_0 は定数項，$\lambda \geq 0$ は縮小度合いを制御するパラメータ，p は Radiomics 特徴量の総数を表す．係数 β_j は，この式の 2 次計画問題を解くことで得られる．しかし，この式の最適の λ の値で求まる Radiomics 特徴量を用いたとき判別性能が最も高い値にはならなかった．そこで本研究では，β_j がゼロでない Radiomics 特徴量の数を 2 個から 10 個になるように λ を設定し，判別性能を計算することで最適な Radiomics 特徴量の数を決定した．まず，Radiomics 特徴量を選択するために 10-fold cross varidation を行い，逸脱度の平均値が最小となる λ を求めた．この計算の過程で求まった複数の λ の値を順番に用いたとき，係数 β_j が非ゼロになる特徴量の数が 2 個から 10 個になる λ の値を採用して Radiomics 特徴量の数を決定した．

2.4　EGFR 遺伝子変異の有無の判別

前節で選択した Radiomics 特徴量を入力とした識別器によって，EGFR 遺伝子変異の有無を判別した．本研究では，識別器として，線形判別分析（Linear Discriminant Analysis, LDA）[16, 17]を用いた．LDA は最も古典的な識別器であるが，結果の解釈が容易であるために採用した．LDA は，Radiomics 特徴量空間において，遺伝子変異ありと遺伝子変異なしの 2 群の分散を同じであると仮定したとき，変異ありと変異なしの 2 群を最も正しく判別する超平面を見つける手法である．LDA の判別得点を用いれば，Radiomics 特徴量空間において各症例が超平面の上か下のどちら側に位置するかを容易に知ることができる．Radiomics 特徴量が病変の表現型を表していることを考慮すれば，判別得点が近い値であるならば，それらの病変の表現型は類似していることを意味する．なお，本研究では，LDA の学習と評価には，Leave-one-out 法[17]を用いた．また，LDA の性能の評価には，シカゴ大学で開発された LABLOC アルゴリズム[18]による ROC 解析を採用し,ROC 曲線以下の面積（AUC）を判別性能として用いた．

2.5　EGFR遺伝子変異ありの遺伝子発現パターンの可視化

本研究では，17302 次元の遺伝子データの分布を 2 次元に投影するために t-SNE[19]を用いた．このような次元削減法には，主成分分析や多次元尺度構成法などがある．しかし，主成分分析や多次元尺度構成法などの線形的な次元削減法では，高次元空間上でデータ分布が非線形な構造を持つとき，類似したデータを低次元空間で近くに表示することが困難であることが知られている．これに対して t-SNE は，高次元空間上でのデータ間の距離が，低次元空間上でのデータ間距離になるべく合致するように変換を行う非線形の次元削減法である．よって，次元削減された後の 2 次元の散布図の状態が高次元空間のデータの分布状態と同じと考えることができる．本研究では，EGFR 遺伝子変異ありの 18 症例の 17302 個の遺伝子データの平均が 0, 分散が 1 になるように正規化処理を加えたものを t-SNE の入力データとして用いた．もし，t-SNE によって変換され

た 2 次元空間において，18 症例が異なる群を形成して分布するならば，同じ EGFR 遺伝子変異ありの肺がんであったとしても，それらは遺伝子発現パターンの異なる群であることを意味する．さらに本研究では，前節の LDA の判別得点の高い群と低い群の 2 群に分類した情報を遺伝子発現パターンの散布図に付加した．これによって，肺がんの遺伝子発現パターンと表現型の関係性を分析することが可能である．

3. 実験結果

Fig.2は，Lasso によって選択された Radiomics 特徴量の数と LDA による判別性能の関係を示す．10 個の Radiomics 特徴量を入力とした LDA の判別性能が最も高かったため，本研究では Radiomics 特徴量の数を 10 個に固定して以下の実験を行った．Table 1 に，選択された 10 個の Radiomics 特徴量を示す．形状に関する特徴量が 5 つ，濃度ヒストグラムに関する特徴量が 1 つ，テクスチャに関する特徴量が 3 つ，解像度に関する特徴量が 1 つ選択された．これらの特徴量の詳細は参考文献を参照されたい[12].

Fig.3に，EGFR 遺伝子変異ありと遺伝子変異なしの合計 119 症例の LDA の出力値をヒストグラムで表示したものを示す．この実験結果で注目すべき点は，EGFR 遺伝子変異ありの症例が 2 つの群に分かれていることである．LDA の出力値は，Radiomics 特徴量空間での判別境界からの距離を表しているから，この結果は同じ EGFR 遺伝子変異ありでも腫瘍の表現型に違いがある群が存在することを示している．以下では，これらの群を Type 1（LDA の値が大きい群）と Type 2（LDA の値が小さい群）として表現する．Fig.4に，Type 1 で LDA の値が大きいものから 6 症例，Type 2 で LDA の値が小さいものから 6 症例を選択した画像を示す．Type 1 と Type 2 で腫瘍内の表現型に違いがあることが理解できる．

Fig.5は，EGFR 遺伝子変異ありの症例で，RNA-Seq のデータが存在した 18 症例の遺伝子発現量を t-SNE を用いて 2 次元で表示した結果を示す．同じ EGFR 遺伝子変異ありの症例でも，遺伝子発現パターンが異なる 2 つの群に分離されているのが分かる．さらに，画像検査による表現型では，Type 1 が左上の遺伝子発現パターンを，Type 2 が右下の遺伝子発現パターンになる傾向があることも明ら

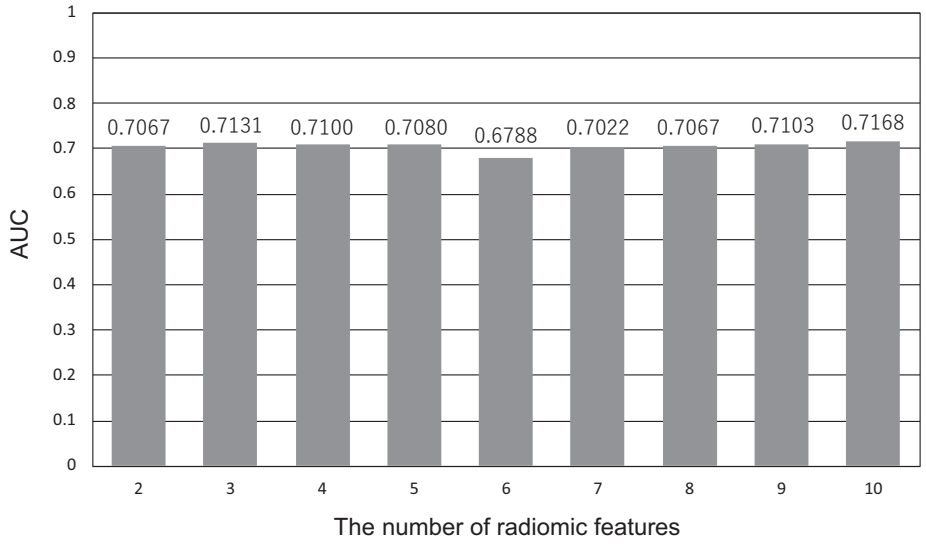

Fig. 2 Relations between the number of radiomic features and AUC obtained by LDA.

Table 1 Selected 10 radiomic features by Lasso.

	Feature	Category	Description
#1	GeoX	Shape	Horizontal coordinate of gravity center
#2	GeoXYo	Shape	Gravity center to inscribed circle center horizontal distance
#3	GeoW6	Shape	Profile specific perimeter squared / 4π*profile area, number of the object pixels
#4	GeoW9	Shape	Area of the circumscribing rectangle of minimal area / Area, number of the object pixels
#5	GeoW13	Shape	Maximal diameter / Area, number of the object pixels
#6	Perc.50%	Histogram	50% percentile
#7	S(2,2) Correlat	Texture	Correlation (S(2,2) is the between-pixels distance)
#8	S(3,0) Correlat	Texture	Correlation (S(3,0) is the between-pixels distance)
#9	S(4,4) Correlat	Texture	Correlation (S(4,4) is the between-pixels distance)
#10	WavEnHL_s-2	Resolution	Wavelet energy (frequency band: HL, scale: 2^{nd} (of 4^{th}))

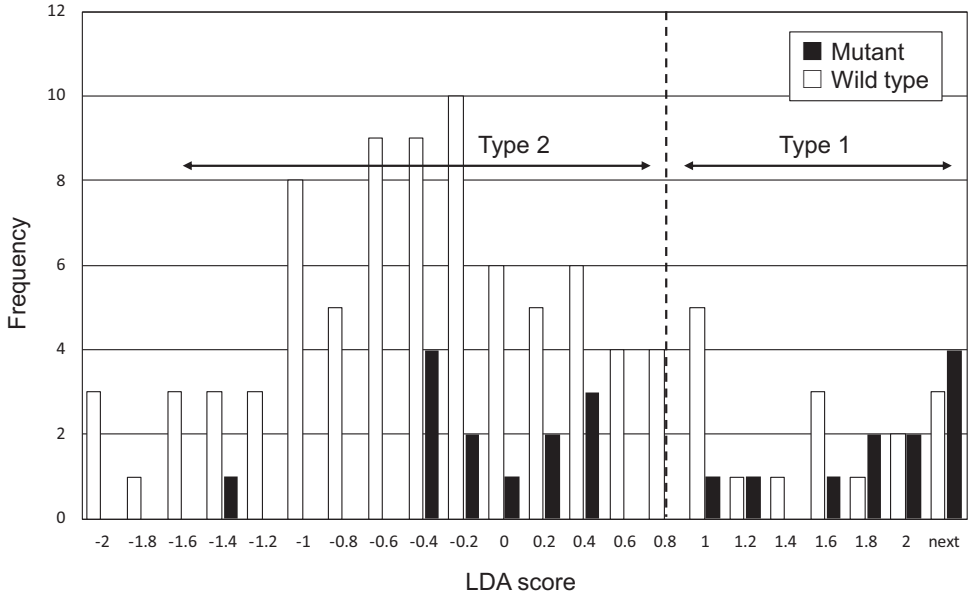

Fig. 3 Histogram of LDA score. The cases with EGFR mutation were classified type 1 and type 2 by using LDA score.

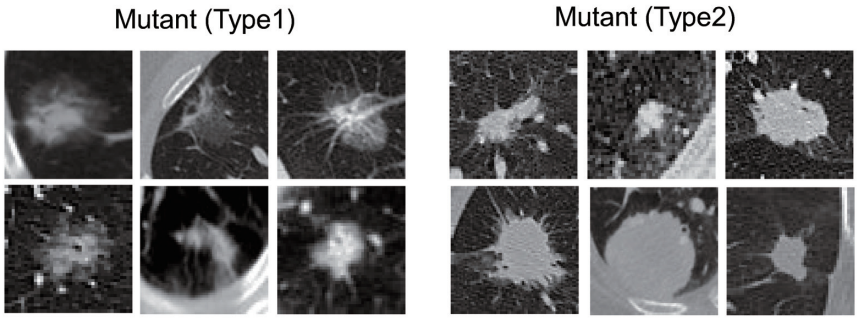

Fig. 4 Examples of type 1 and type 2. These are cases with EGFR gene mutation.

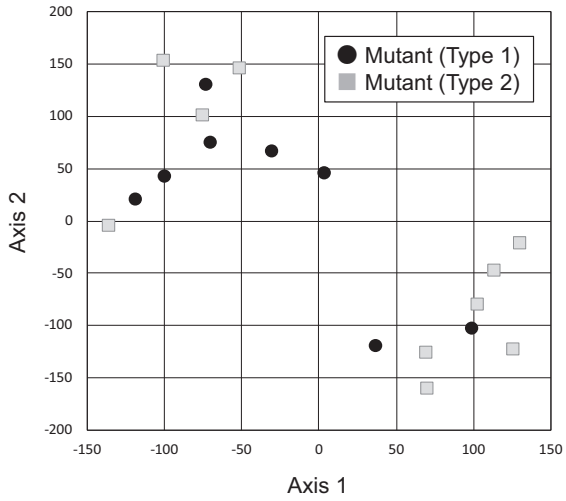

Fig. 5 Output of t-SNE when gene expression levels were used as input data. Type 1 and type 2 represent imaging phenotypes of tumors.

かになった．Type 1 と Type 2 で左上の群と右下の群に分類して，2×2 の分割表を作成し，独立性の検定を行った．しかし，p 値は 0.137 であって有意差はなかった．

4. 考　察

Fig.3の結果から，10 個の Radiomics 特徴量を選択した場合と 3 個の Radiomics 特徴量を選択した場合に，AUC 値に大きな差が無かった．Fig.6に，10 個の Radiomics 特徴量を用いた場合と 3 個の Radiomics 特徴量を用いた場合の LDA の出力値の関係を示す．選択された 3 個の Radiomics 特徴量は，Perc.50%,S(3,0)Correlat,S(4,4)Correlat であり，Table 1 に示す 10 個の Radiomics 特徴量に含まれる．LDA の出力値の相関係数は 0.81 であった．3 個の Radiomics 特徴量を用いた場合も 10 個の Radiomics 特徴量を用いた場合と同様の傾向がある．そこで本論文では，よ

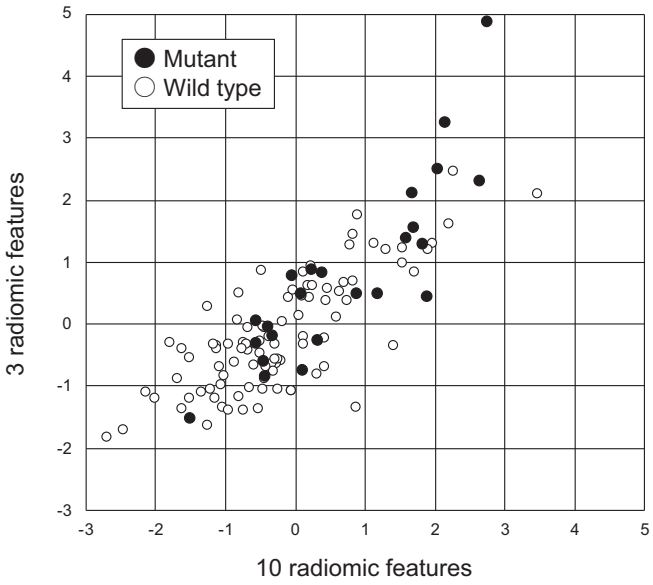

Fig. 6 Relationship between LDA outputs used 10 radiomics features and LDA outputs used 3 radiomic features.

り判別性能が高い 10 個の Radiomics 特徴量を用いた結果について，以下で考察する．

Fig.3の結果は，同じ EGFR 遺伝子変異ありの肺がんでも，その表現型が EGFR 遺伝子変異なしの群と明確に異なる群（Type　1）と EGFR 遺伝子変異なしの群に近いもの（Type　2）が存在することを示している．また，Fig.5の結果から，同じ EGFR 遺伝子変異ありの症例でも Type　1 と Type　2 では，遺伝子発現パターンも異なる傾向があることが示唆された. 細胞のがん化は, 遺伝子変異の蓄積によって進むため，その変異の蓄積パターンは症例ごとに異なるのが一般的と考えられる．さらに，その変異パターンの違いが表現型に現れているとも考えられる．したがって，腫瘍の表現型から EGFR 遺伝子変異の有無を推定する Radiomics 研究においては，EGFR 遺伝子変異ありと遺伝子変異なしの 2 群の分類問題として取り扱うのではなく，他の遺伝子変異パターンも考慮した分類問題として検討した方が適切であると考えられる．

EGFR 遺伝子変異は，肺がんにおいて高頻度で発見されるため，EGFR 阻害剤が開発されている．しかし，EGFR 遺伝子変異ありの群でも遺伝子発現量のパターンが異なる群が存在することを考慮すれば，EGFR 阻害剤のみで奏効する症例とそうでない症例が存在する可能性が高いことは容易に想像できる．本研究によって，非侵襲に腫瘍の表現型からある遺伝子発現パターンを伴う EGFR 遺伝子変異ありの症例（Type 1）が容易に検出できることを示したが，このことを至適治療法の選択を支援するシステムの開発に繋げるには，Type 1 の肺がん患者に対して，EGFR 阻害剤を用いたときの奏効率を調査する必要がある．もし，Type 1 の肺がんが EGFR 阻害剤の効果が高いならば，EGFR 遺伝子変異ありで EGFR 阻害剤の効果が期待できる群，EGFR 遺伝子変異ありで EGFR 阻害剤の効果が期待できない群，EGFR 遺伝子変異なしの群の 3 つの群の分類問題として取り扱えば良い．本研究で用いたデータベースには，EGFR 阻害剤の奏効率に関する情報が無かったため，このことは今後の検討課題であると考えられる．

本研究のリミテーションは，RNA-Seq のデータが存在する EGFR 遺伝子変異ありの症例が 18 症例しかないことである．また，腫瘍内不均一性の問題から，RNA-Seq の遺伝子パターンが腫瘍全体の性質を捉えておらず偏っている可能性も否定できない．今後，多くの症例を収集して実験結果の再検証をする必要があると考えられる．画像に関するリミテーションは，複数の施設から収集した CT 画像を用いているため，異なる撮影条件による Radiomics 特徴量のバラツキが実験結果に影響を及ぼしている可能性が否定できないことである．この点に関して詳細に検討する必要があると考えられる．また，本研究では CT 画像から 1 枚のスライスを選択して 2 次元の Radiomics 特徴量を計測したが，3 次元の Radiomics 特徴量を計測すれば判別性能が向上する可能性もある．さらに，腫瘍領域のマーキングの違いの影響が推定精度に影響を及ぼす可能性も考えられるため，今後の検討課題としたい．

5. 結　語

同じ EGFR 遺伝子変異ありの肺がんでも，遺伝子発現パターンが異なることが明らかになった．また，肺がんの Radiomics 特徴量を用いれば，ある遺伝子発現パターンを持つ EGFR 遺伝子変異ありの肺がんを比較的容易に検出できることも示した．画像検査によって，肺がんの遺伝型を推定する Radiomics 研究においては，EGFR 遺伝子変異ありと遺伝子変異なしの 2 群を分類する問題として考えるのではなく，EGFR 阻害剤の奏効率と関係する肺がんの遺伝子発現パターンも考慮した多群の分類問題として検討した方が良いことが示唆された．

謝　辞

本研究の一部は，JSPS 科研費基盤研究 C（課題番号 21K12707）にて行われた．

参考文献

[1] Weinberg　RA（著），武藤誠（翻訳），青木正博（翻訳）：ワインバーグがんの生物学，第 2 版，南江堂，2017.
[2] 国立がん研究センター中央病院呼吸器内科：最先端治療肺がん，法研，2016.
[3] Jia TY, Xiong JF, Li XY, et al.: Identifying EGFR mutations in lung adenocarcinoma by noninvasive imaging using radiomics features and random forest modeling, European Radiology, 29(9) : 4742-4750, 2019.

[4] Tu W, Sun G, Fan L, et al.: Radiomics signature: A potential and incremental predictor for EGFR mutation status in NSCLC patients, comparison with CT morphology, Lung cancer, 132 : 28-35, 2019.
[5] Nair JKR, Saeed UA, McDougall CC, et al.: Radiogenomic models using machine learning techniques to predict EGFR mutations in non-small cell lung cancer, Can Assoc Radiol J, 72(1) : 109-119, 2021.
[6] Li Y, Lu L, Xiao M, et al.: CT slice thickness and convolution kernel affect performance of a radiomic model for predicting EGFR status in Non-small cell lung cancer: A preliminary study, Scientific Reports 8(1) : 17913, 2018.
[7] Hong D, Xu K, Zhang L, et al.: Radiomics signature as a predictive factor for EGFR mutations in advanced lung adenocarcinoma, Front Oncol, 10 : 28, 2020.
[8] Rossi G, Barabino E, Fedeli A, et al.: Radiomic detection of EGFR mutations in NSCLC, Cancer Res, 81 (3) : 724-731, 2021.
[9] Wu S, Shen G, Mao J, et al.: CT radiomics in predicting EGFR mutation in non-small cell lung cancer: A single institutional study, Front Oncol 10 : 542957, 2020.
[10] Li XY, Xiong JF, Jia TY, et al.: Detection of epithelial growth factor receptor (EGFR) mutations on CT images of patients with lung adenocarcinoma using radiomics and /or multi-level residual convolutionary neural networks, J Thorac Dis, 10(12) : 6624-6635, 2018.
[11] https : / / wiki. cancerimagingarchive. net / display / Public / NSCLC+Radiogenomics
[12] MaZda, http : //eletel.eu/mazda
[13] Szczypinnski PM, Strzelecki M, Materka A, et al.: MaZda-a software package for image texture analysis, Comput Methods Programs Biomed., 94(1), 66-76, 2009.
[14] Strzelecki M, Szczypinski P, Materka A, et al.: A software tool for automatic classification and segmentation of 2D/3D medical images, Nucl Instruments Methods Phys Res., 702, 137-140, 2013.
[15] Hastie T, Tibshirani R, Friedman J.: The elements of statistical learning, data minimg, inference and prediction, second edition, Springer New York, 2009.
[16] Theodoridis S, Koutroumbas K.: Pattern recognition. Academic Press, London, 1999.
[17] Duda RO, Hart PE, Stork DG.: Pattern Classification. New York : John Wiley & Sons, 2001.
[18] Metz CE: Sonic practical issues of experimental design and data analysis in radiological ROC studies, Invest Radiol., 24(3), 234-245, 1989.
[19] Maaten LVD, Hinton G: Visualizing data using t-SNE, Journal of machine learning research, 9 : 2579-2605, 2008.