Diagnostic utility of molecular profiling to distinguish multiple primary lung cancer from intrapulmonary metastasis

Yeon Seung Chung

Department of Medicine

The Graduate School, Yonsei University

Diagnostic utility of molecular profiling to distinguish multiple primary lung cancer from intrapulmonary metastasis

Directed by Professor Hyo Sup Shim

The Doctoral Dissertation
submitted to the Department of Medicine,
the Graduate School of Yonsei University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Medical Science

Yeon Seung Chung

December 2023

This certifies that the Doctoral Dissertation of
Yeon Seung Chung is approved.

Thesis Supervisor : Hyo Sup Shim

Thesis Committee Member#1 : Chang Young Lee

Thesis Committee Member#2 : Sunhee Chang

Thesis Committee Member#3: Hye Ryun Kim

Thesis Committee Member#4: Sangwoo Kim

The Graduate School
Yonsei University

December 2023

ACKNOWLEDGEMENTS

First of all, I would like to express my sincere gratitude to Professor Hyo Sup Shim for his guidance throughout this entire study. His assistance has been paramount in helping me complete my doctoral course and fostering my interest in research and practical work.

Also, I extend my infinite gratitude to Mr. Jeongsoo Won and Professor Sangwoo Kim from the Department of Biomedical Systems Informatics. Their insights and teachings on bioinformatic analysis, an area I was previously unfamiliar with, as well as their assistance in the model development, have made the completion of this thesis possible.

And I would like to express my deepest gratitude to Professor Chang Young Lee, Professor Sunhee Chang, and Professor Hye Ryun Kim for graciously agreeing to review this thesis. Their invaluable insights and broad perspective have illuminated aspects I had not considered, and their guidance has been instrumental in the completion of this thesis.

Lastly, I would like to express my appreciation to my family, who always share my everyday life.

<TABLE OF CONTENTS>

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Diagnostic utility of molecular profiling to distinguish multiple primary lung cancer from intrapulmonary metastasis

Yeon Seung Chung

Department of Medicine
The Graduate School, Yonsei University

(Directed by Professor Hyo Sup Shim)

Introduction: In multiple lung cancer, distinguishing between multiple primary lung cancer (MPLC) and intrapulmonary metastasis (IPM) critically influences clinical prognosis prediction and therapeutic decision-making. Despite various approaches proposed, leveraging histological assessments and molecular status, the task remains challenging.

Methods: We introduced the MeTel (Metastasis-Teller) algorithm, a Bayesian probabilistic model designed to determine MPLC or IPM based on the molecular profile of the tumor. Six datasets from previous studies, encompassing 279 tumor pair (75 IPM and 204 MPLC) with diverse sizes of gene panel (22 genes to whole exome sequencing) were compiled to compare the accuracy of MeTel against other previously published algorithms. Equivocal cases from our institution were further re-classified using MeTel, with results validated using next-generation sequencing or whole-exome sequencing.

Results: MeTel exhibited superior performance with 97.5% accuracy across all six datasets, outpacing other algorithms which ranged from 82.08% to 95.70%. Importantly, its accuracy remained consistent regardless of gene panel size. In our institution's evaluation of the 12 equivocal cases, four cases showed discordant result with histologic criteria; subsequent validation favored MeTel's classifications.

Conclusion: MeTel demonstrates reliable accuracy in classifying multiple lung cancer and holds significant potential for clinical application in the future.

Key words : multiple primary lung cancer, intrapulmonary metastasis, multifocal lung cancer, non-small cell lung cancer, Bayesian probabilistic model

Diagnostic utility of molecular profiling to distinguish multiple primary lung cancer from intrapulmonary metastasis

Yeon Seung Chung

Department of Medicine
The Graduate School, Yonsei University

(Directed by Professor Hyo Sup Shim)

I. INTRODUCTION

Lung cancer is the second most commonly diagnosed cancer and the leading cause of cancer death worldwide, with up to 75% of recurrence rates in non-small cell lung carcinoma(NSCLC)[1], and 90% in small cell lung carcinoma(SCLC)[2], within 2 years after surgical resection. The prognosis of lung cancer are significantly determined by the TNM staging system[3], so accurate TNM staging is pivotal for both prognostic assessment and therapeutic planning in lung cancer patients.

One of the factors that contribute to determining the TNM stage of lung cancer is its multifocality. The presentation of multiple tumors in lung cancer cases is not uncommon, reported in 0.2 to 20% of cases.[4] The TNM stage of multiple lung cancers differ widely, depending on whether they are intrapulmonary metastases(IPMs) or multiple primary lung cancers(MPLCs). In the 8th edition of the AJCC staging manuals, multiple tumor nodules categorized as IPM are classified as pT3 if they are located in the same lobe, pT4 if in a different but ipsilateral lobe, and M1a if in the contralateral lobe. Meanwhile, tumor nodules categorized as SPLC are staged according to the T stage for each tumor independently.[5] This distinction is crucial for deciding whether patients should have surgical resection or non-invasive treatments like chemotherapy or radiotherapy. For example, in bilateral lung cancer cases, surgical resection may be an option for MPLC but only palliative care with

chemoradiation therapy is available for contralateral IPM.[6]

Differential diagnosis between MPLC and IPM is crucial yet challenging in clinical practice, prompting several efforts to distinguish between the two.[7-10] The first criteria for distinguishing MPLC and IPM were proposed by Martini and Melamed in 1975.[7] According to their criteria, MPLC is defined based on whether the multiple lung cancers have differing histologies. If the histologies are identical, MPLC is defined: each tumor shows carcinoma in situ, with an absence of lymph node or extrapulmonary metastases, and, if the tumors occur metachronously, a time interval of at least two years exist. Subsequently, the American College of Chest Physicians(ACCP) proposed a modified criterion in 2007 that extended the time interval to four years.[8] Additionally, Girard *et al*. introduced further diagnostic criteria, including histologic subtype as well as cytologic and stromal features.[9] These criteria focus on the tumor's clinicopathologic features, which simplifies their clinical application. However, the lack of high-level evidence, such as molecular profiling of the tumor, complicates the clinical decision-making process for accurately defining the true nature of the tumors.

Recently, researchers have actively been trying to use genomic technology to classify SPLC and IPM. Driver mutation statuses such as EGFR, ALK, and ROS1 are now commonly evaluated for lung adenocarcinoma, and several studies have attempted to use these driver mutation statuses for the classification of SPLC and IPM, given that tumors from the same origin share driver mutations, though the opposite isn't always true.[11, 12] More recently, large-scale gene comparisons using next-generation sequencing(NGS) have been employed to distinguish between SPLC and IPM, extending beyond the scope of just driver oncogenes. Nicholson *et al* developed a strategy (named Cochin) that uses both clinical information (time interval) and the number of shared non-driver mutations, wherein any shared rare mutation is judged to be IPM[13]. Chang *et al* used the same strategy on the number of shared mutation, with further utilizing histological information (named MSK)[14]. Wang *et al* developed the first genome-only classification, HAPLOX[15], allowing up to two non-driver mutations

for MPLC. The rationale for using NGS sequencing-based discriminatory criteria is clear - the more mutations two tumors share, the more likely they are to be IPMs of the same origin. However, classification based on the count of shared mutations has its limitations, as the cutoff is inherently arbitrary. Additionally, the reliability of this criterion can vary depending on the size of the gene panel, as a larger panel increases the chance of coincidental matching mutations between independent tumors.

In this study, we propose a new algorithm for differentiating MPLC and IPM based on a Bayesian probability model, which we have named 'Metastasis-Teller' as ‘MeTel’. MeTel calculates the probability ratio for MPLC and IPM in a consistent manner, regardless of panel size, and does not require a specific cutoff. The probability for MPLC or IPM reflects the variant allele frequency(VAF) and occurrence rate of the mutations detected from each tumor, offering a more quantitative and objective criterion compared to the empirical interpretation of the previous studies.

We assessed MeTel's performance by comparing it against other NGS-based algorithms (Cochin, MSK, and HAPLOX), using six multiple lung cancer cohorts totaling 279 samples, with gene panel sizes ranging from 22 genes to whole-exome sequencing. Additionally, we gathered ambiguous MPLC or IPM cases under histological criteria within our institution and reclassified them using MeTel.

II. MATERIALS AND METHODS

1. MeTel Algorithm

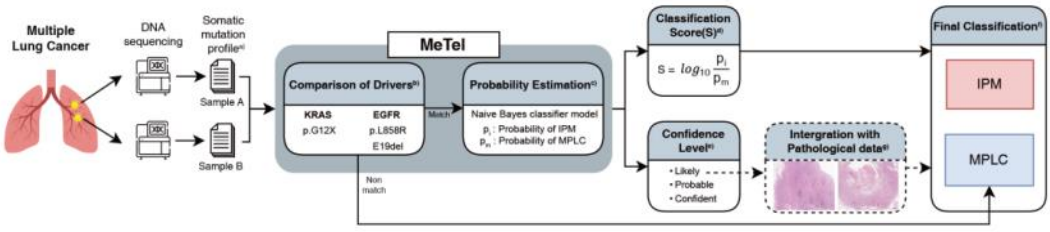


Figure 1. Overview of MeTel algorithm. a) MeTel takes in input somatic mutation (with variant allele frequency) profile from DNA sequencing data of multiple lung cancer samples. b) First, MeTel compares driver mutations (*KRAS* p.G12X, *EGFR* p.L858R and E19del). If there are different drivers, they are classified as multiple primary lung cancer (MPLC), and if the driver matched, move on to step c). c) MeTel estimates probability of intrapulmonary metastasis (IPM) (p_i) and MPLC (p_m). d) MeTel outputs classfication score (S), the log-scale value of the ratio of p_i, and p_m. e) Confidence level, another output from MeTel. Based on the maximum value of p, and p, 0.8 and 0.95 cutoff represent one of the three: Likely, Probable and Confident. f) Final classification IPM or MPLC: If S > 0, samples classified as IPM or if S < 0, MPLC. g) The process of combining with histopathology data with Metel's results (only with 'Likely' confidence level).

We introduce MeTel, an innovative algorithm designed to classify IPM and MPLC using molecular information from multiple lung cancer samples obtained from the same patient. MeTel takes genomic profiles, including integrated somatic mutations and variant allele frequencies (VAFs) of tumor samples, as input (Fig. 1a).

The primary criterion employed by MeTel for classification is the presence of matching driver mutations (Fig. 1b). If the tumor pair exhibits KRAS (p.G12X) or EGFR (p.L858R or E19del) mutations, known to be the most common driver mutations in NSCLC, as the sole driver mutations, it is classified as MPLC. However, if the tumors cannot be classified based on this criterion, MeTel proceeds to estimate the probabilities of IPM and MPLC using a Naive Bayes Classifier (Fig. 1c).

We assume that the occurrence of each mutation is independent and define the observation value D as the union of observed somatic mutations and VAFs in the early

and late tumor samples, denoted by A and B. We establish the following equation based on the Naive Bayes Classifier for $D = \{v_1, v_2, \cdots, v_n\}$, and p(IPM) and p(MPLC), which are the priors of IPM and MPLC :

$$P(IPM|D) = \prod_{i=1}^{n} P(v_i|IPM)p(IPM) \quad (1)$$

$$P(MPLC|D) = \prod_{i=1}^{n} P(v_i|MPLC)p(MPLC) \quad (2)$$

The status of the variable v_i, belonging to D, can manifest in one of three forms: (VAF in A, VAF in B), (VAF in A, not detected in B), and (not detected in A, VAF in B). To account for these possibilities, we calculated six distinct likelihood values applicable under the conditions of IPM and MPLC.

$$P(VAF_A, VAF_B|IPM) = p * \{m + (1 - m) * p\} \quad (3)$$

$$P(VAF_A, \text{Not in B}|IPM) = p * \{(1 - m) * (1 - p)\} \quad (4)$$

$$P(\text{Not in A}, VAF_B|IPM) = (1 - p) * p \quad (5)$$

$$P(VAF_A, VAF_B|MPLC) = p * p \quad (6)$$

$$P(VAF_A, \text{Not in B}|MPLC) = p * (1 - p) \quad (7)$$

$$P(\text{Not in A}, VAF_B|MPLC) = (1 - p) * p \quad (8)$$

In the equations, p represents the probability of the corresponding somatic variant occurring accidentally in lung cancer. This value is determined based on the frequency with which the mutation was reported in the NSCLC dataset of the GENIE database (v13) from cBioPortal. It is calculated as $p = n_{case}$ (reported case in primary NSCLC) / n_{all} (all case in primary NSCLC). If the mutation is not present in the database, we set p to 10^{-6}, which is generally the probability of accidental mutation occurrence.

The variable m represents the probability of the variant being transmitted during tumor

metastasis and is proportional to the VAF, which denotes the ratio of alleles with the corresponding mutations. Assuming no copy number variation (CNV) or loss of heterozygosity (LOH) has occurred, m is set to 2 * VAF for cells with two alleles. If the VAF is 0.5 or greater ($m \geq 1$), m is set to $1 - \varepsilon$ to avoid divergence within the equation. In the absence of available VAF information, the default value is set to 0.3 which is the mean VAF of mutations reported in the GENIE database used to determine the value of p. Alternatively, the expected average VAFs can be manually inputted.

Equations (3) to (5) represent the likelihoods for the IPM condition, while equations (6) to (8) represent the MPLC condition. In Equation (3), the mutation is observed in both A and B, and it can occur through two possible pathways: either the mutation arises in A and is subsequently transmitted to B through metastasis, or the mutation arises independently in both A and B. Equation (4) describes the situation where the mutation is observed in A only, indicating that the mutation occurred in A but was not transferred to B through metastasis and did not occur independently in B. Equation (5) pertains to the scenario where the mutation is only observed in B, indicating that the mutation did not arise in A but appeared in B after metastasis. For the MPLC condition, where A and B are independent, equations (6) to (8) are computed by multiplying each probability.

We set the prior probabilities as p(IPM) = 0.28 and p(MPLC) = 0.72 based on the respective ratio of IPM and MPLC observed in 975 samples across seven existing study datasets[13-19].

Using the likelihoods (3) to (8) and the priors, we compute equations (1) and (2) to estimate the probabilities of IPM (p_i) and MPLC (p_m) : $p_i = (1) / \{(1) + (2)\}$ and $p_m = (2) / \{(1) + (2)\}$.

After the estimation process, MeTel provides two values: the classification score (S) (Fig. 1d) and corresponding confidence level (Fig. 1e). The classification score (S) is the logarithmic scale value of the ratio of two probabilities: $S = \log_{10} \{p_i / p_m\}$.

If the sequence of the two samples is unknown, we calculate two S values based on the sequence and choose the case with the larger absolute value (|S|). The confidence level reflects the reliability of the S value and is categorized into three classes: Likely, Probable, and Confident, based on the difference between the estimated p_i and p_m from the previous step. MeTel assigns the confidence level using MAX (p_i, p_m) according to the criteria 0.8 and 0.95 (Supplementary table 1). Consequently, MeTel distinguishes whether the relationship between tumors is IPM or MPLC: $S > 0$ suggests IPM, while $S < 0$ suggests MPLC (Fig. 1f).

Following the MeTel classification, an optional process is available for samples with a 'Likely' confidence level (Fig. 1g). The 'Likely' condition indicates relatively low reliability in the algorithm's classification call, and we suggest an additional process that combines histopathological data to achieve a more sophisticated classification. In this step, we demonstrate the possibility of modifying the original results when the previous classification differs from the new classification obtained through the optional process.

2. Test Dataset

Table 1. Summary of test dataset

Panel size of dataset (number of gene)	Number of Pairs (IPM)	Number of Pairs (MPLC)	Clinical information[1]	Reference
22	33	76	△	[13]
189	14	0	O	[20]
468	25	51	O	[14]
520	3	22	O	[18]
605	7	44	X	[15]
WES	0	11	O	[21]

Total 6 datasets with 279 paired samples. (75 IPM/204 MPLC)

[1]O: All cases were available for histologic classification (MPLC or IPM). △: Only 30 pairs were available for histologic classification, which showed discordant results with molecular classification in the study. X: Histologic classification was not available, but all cases were synchronous adenocarcinomas.
IPM: intrapulmonary metastasis, MPLC: multiple primary lung cancer, WES: whole-exome sequencing.

To evaluate the performance of algorithms, we tested four different methods, including previous approaches such as Cochin, MSK, and HAPLOX (Supplementary table 2)[13-15] on six distinct datasets (Table 1)[13-15, 18, 20, 21]. The test dataset comprised somatic mutation profiles and clinical data of each sample from existing studies. Tumor pair with no genetic alteration (all-wild type) was excluded during the collecting process.

To determine the diagnosis of each paired sample, we relied on the final classified results from previous studies. The test dataset consisted of a total of 279 pairs, with 75 pairs diagnosed with IPM and 204 pairs diagnosed with MPLC. The panel sequencing data utilized in the dataset varied in the number of genes covered, including 22, 189, 468, 520, 605, and whole-exome sequencing(WES).

During the evaluation of algorithm performance, limited clinical data were available from the studies. In the absence of specific information, we applied a default VAF value of 0.3. When evaluating the performance of previous algorithms, we compared the results obtained both with and without training datasets specific to each algorithm.

3. Sample Background for in-house patients

To evaluate the scalability and potential applicability of MeTel, we gathered NSCLC cases from our institution as a resource for comprehensive molecular investigation. We reviewed the medical charts of NSCLC patients who underwent surgical resection at Yonsei University Severance Hospital (Seoul, Korea) during 2006 to 2020. During the study period, we identified 4595 patients, out of which 493 patients (10.7%) were pathologically confirmed to have multiple tumor nodules. We excluded 60 patients who did not achieve complete tumor resection at the time of their final surgical intervention. Consequently, the total number of eligible patients available for this study culminated in 433.

From the eligible patients, Clinical data including age, sex, smoking history, location of tumor, nodal status were obtained from medical records. Disease-free survival was defined as the time from the last resection to the point of recurrence detected in the follow-up examination (imaging or biopsy). The histology of the tumors was reviewed by pathologists (YSC and HSS) using the comprehensive histologic assessment (CHA) proposed by Girard *et al*.[9] Additionally, if a driver gene mutation test was performed

at the time of diagnosis, the results of the tests were also collated.

4. Statistical analysis

For a descriptive study of in-house patients according to classification (MPLC or IPM), categorical variables were presented as frequencies and percentages. Continuous variables were described using medians and ranges. Variables in each classification were compared using Mann-Whitney U test and $\chi 2$ test. Disease-free survival curves for each classification were estimated using the Kaplan-Meier method and compared with the log-rank test. All results were considered significant when the significance tests indicated a two-sided p value of less than 0.05. Statistical analysis was performed using SPSS Version 26.0 statistical software (IBM SPSS Statistics for Windows, Version 26.0. Armonk, NY: IBM Corp.).

5. In-house patient selection

Table 2. Summary of clinical characteristics of the 12 in-house patients.

Patient characteristics (N=12)	Value
Sex, n (%)	
Male	4 (33.3)
Female	8 (66.7)
Mean age at first resection, y (range)	63.7 (45-74)
Smoking status, n (%)	
Current/Ex-smoker	2 (16.7)
Nonsmoker	10 (83.3)
Mean pack-year of smoker (range)	27.5 (25-30)
Synchronicity, n (%)	
Synchronous	2 (16.7)
Metachronous	10 (83.3)
Median time interval for metachronous case, m (range)	64.6 (13.3-129.8)
Distribution of tumors, n (%)	
Ipsilateral (same lobe)	1 (8.3)
Ipsilateral (different lobe)	5 (41.7)
Contralateral	6 (50.0)

After a comprehensive review of the cases, tumor pairs showing atypical features under traditional histological criteria were ultimately selected for MeTel application. Initially, 12 in-house patients with 27 tumors were chosen, but two tumors were excluded due to low sequencing quality, resulting in a total of 25 tumors being included in the study for the following reasons (Table 2 and Supplementary Table 4):

(1) Similar histology with a long time interval: Traditional histologic criteria define a time interval cutoff for the determination of MPLC - 2 years according to Martini and

Melamed criteria[7], and 4 years according to ACCP[8]. In the Cochin method[13], 5-year cutoff is suggested for MPLC identification. We identified five patients who presented multiple tumors with similar histology and shared driver gene mutation, but had a time interval exceeding five years (patient 1-5 and 7-8).

(2) Synchronous multiple squamous cell carcinoma: Determining the origin of squamous cell carcinoma (either MPLC or IPM) is challenging due to the lack of a specific driver gene and histologic diversity[22, 23], compared to adenocarcinoma. We identified a pair of synchronous squamous cell carcinomas with similar histology and no precursor lesion, implying an IPM diagnosis, but with a favorable prognosis. Both tumors were found on contralateral sides, which significantly increased the tumor stage when considered as IPM. However, the patient survived without recurrence for over 71 months (patient 6).

(3) Similar histology with discordant driver gene status: One patient exhibited recurrent adenocarcinoma in the same lobe after a year, presenting a similar histologic subtype (acinar predominant with papillary and micropapillary components) devoid of any in situ lesion. Nevertheless, the subsequent tumor acquired an EGFR exon 19 deletion, unlike the previous tumor, which was of EGFR wild type. This introduced a discordance between the clinicohistologic indication for IPM and its driver gene status (patient 9).

(4) Ambiguous histology with identical driver gene status: Some patients exhibited ambiguous histologic features across multiple tumors with the same driver gene status. They shared a similar histologic subtype components, but the proportion of the subtype had changed, some including the dominant histologic pattern, leading to an MPLC identification. However, as previously noted, they share the same histologic patterns and driver gene mutation, which does not eliminate the possibility of IPM. (patient 10-12).

6. Genomic data Profiling

To obtain reliable somatic mutations for the classification algorithm, we performed targeted sequencing using the TSO500 DNA workflow[24] for 12 samples from 6 patients. We then filtered out variables with a VAF outside the range of 0.05 to 0.5 and applied germline filtering using its own population database in the TSO500 pipeline. To ensure the accuracy of the results, each somatic mutation was manually checked using IGV[25].

For a more comprehensive analysis, we also performed whole-exome sequencing (WES) for 6 patients with a total of 13 samples with a sequencing depth of 200x. The raw reads were aligned to the GRCh38 genome reference using the BWA-MEM aligner (v0.7.17-r1188)[26], and pre-processing was completed by applying MarkDuplicates, FixMateInformation, BaseRecalibrator, and ApplyBQSQ, which are included in the GATK Best Practices (v4.0.1.1)[27].

Somatic mutations were called using the GATK Mutect2 (v.4.0.1.1) tumor-only mode with the --f1r2-tar-gz argument to remove strand orientation bias artifacts. We filtered cross-sample contamination using GetPileupSummaries and CalculateContamination, and further artifact of FFPE filtering was applied using SOBDetector (v1.0.2)[28] to the filtered output of Mutect2. To exclude false calls, we applied filtering based on VAF (<0.05) and the count of alternate allele (<5). Additionally, we applied the following cutoffs to remove germline mutations: (1) VAF (>0.5), (2) variants from two germline mutation callers, Strelka2 (v2.9.10)[29] and GATK HaplotypeCaller (under default parameters), (3) dbSNP (>0.01), and (4) gnomAD db (>0.0001)[30]. This rigorous filtering process ensured that only high-quality somatic mutations were included in our analysis, improving the accuracy of the classification algorithm.

III. RESULTS

1. Performance of Algorithm

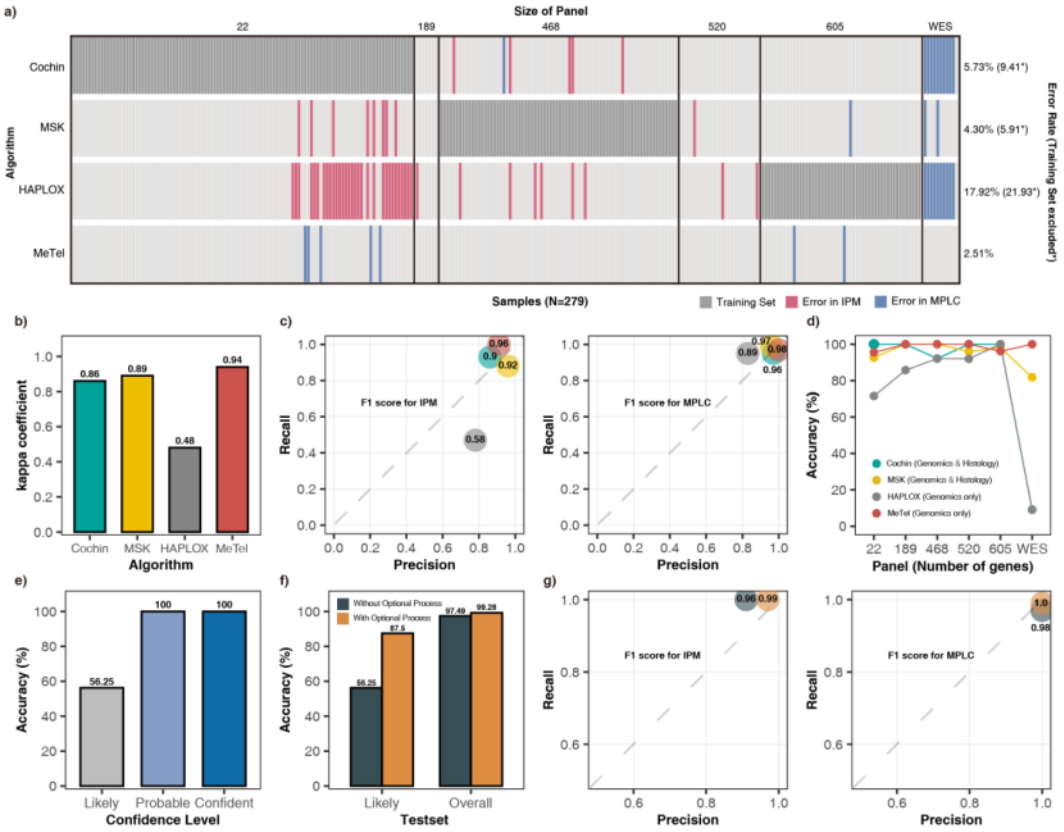


Figure 2. Performance of algorithms. a) Classification results of algorithms for test set (N=279). Dark gray dataset mean the training set of the algorithm. *, error rate except for training dataset. b) The Cohen κ scores by algorithms. c) Precision and Recall for intrapulmonary metastasis (IPM) and multiple primary lung cancer (MPLC). The values represent the F1 scores. d) The accuracy of the algorithms on the dataset of different sizes of panels. e) Accuracy by confidence level. f) Changed accuracy at 'Likely' and overall with optional process. g) Precision and recall before and after optional process.

We compared MeTel with three previous methods (Cochin, MSK, and HAPLOX) for distinguishing between IPM and MPLC using the test dataset. MeTel outperformed all other methods, achieving the lowest error rate of 2.51% for total test set (Fig. 2a and

Supplementary table 5). This error rate was less than 60% of the error rates observed with the other algorithms. MeTel also showed the highest Cohen κ score of 0.94 (Cochin: 0.86, MSK: 0.89, and HAPLOX: 0.48), indicating almost perfect agreement with the correct classification[31] (Fig. 2b and supplementary table 5). The F1 scores for IPM and MPLC were 0.96 and 0.98, respectively, demonstrating the strong predictive power of the MeTel algorithm (Fig. 2c and supplementary table 6). Notably, MeTel which is based solely on genomic data, outperformed algorithms that incorporate clinical data. The outstanding performance of MeTel was further highlighted in the results excluding the training dataset of each algorithm that showed 100% accuracy.

We also evaluated the performance of the methods using different panel sizes, ranging from 22 genes to WES. Each algorithm was optimized for performance on the trained dataset, resulting in performance deviations when applied to other datasets, particularly when there were significant differences in panel size. However, MeTel consistently demonstrated excellent performance across most datasets with only a slight decline in accuracy, even though it was not trained on any of the test sets used to developed the algorithm (95.41% - 100%) (Fig. 2d). MeTel showed exceptional performance even on the large sequencing dataset WES, which had the most significant difference compared to other algorithms (Cochin: 9.09%, MSK: 81.82%, and HAPLOX: 9.09%).

MeTel provides three confidence levels based on the difference between the probabilities of IPM and MPLC. For the total test set (n=279), the number of samples included in each confidence level were as follows: Likely: 16 (5.73%), Probable: 80 (28.67%), and Confident: 183 (65.59%). Errors in MeTel were only observed in ‘Likely’ cases, which had a low number of variants on average (1.29), while all cases categorized as Probable and Confident showed 100% accuracy (Fig. 2e and Supplementary Fig.1) Thus, the confidence level of MeTel provides important evidence for decision-making when integrating clinical data with results of MeTel (optional process).

2. Integration of Optional Pathological Data

While MeTel demonstrated high accuracy using genomic data alone, we proposed an optional process to incorporate histopathological data based on MeTel's output.

In cases where the classification was not clear due to minimal differences in the probabilities between IPM and MPLC, we applied an exceptional rule to adjust the original results of MeTel based on histology. This optional process was performed on 16 'Likely' cases, resulting in an improvement in accuracy. Of the total seven errors, 5 errors belonging to the 22-panel dataset were correctly classified (two were included in the 605-panel dataset without available clinical data). As a result, the accuracy of MeTel for the 'Likely' category improved to 87.5%, and the accuracy for the entire test set was 99.28% (Fig. 2f). Furthermore, the F1 scores for both IPM and MPLC reached 0.99 and 1.0 (Fig. 2g), respectively, indicating that combining MeTel's results with histopathological information allows for nearly perfect prediction.

3. Descriptive analysis of sample background of in-house patients

Table 3. Clinico-histologic characteristics of 433 non-small cell lung cancer patients with multiple tumors at Yonsei University Severance Hospital (2006 to 2020)

	IPM (N=156)	MPLC (N=277)	p-value
Sex, n (%)			**<0.001**
Male	104 (66.7)	143 (51.6)	
Female	52 (33.3)	134 (48.4)	
Median age at first resection (range)	64.0 (32-83)	65.0 (41-84)	**0.027**
Smoking status, n (%)			**<0.001**
Current smoker	51 (32.7)	61 (22.0)	
Ex-smoker	51 (32.7)	63 (22.7)	
Nonsmoker	54 (34.6)	153 (55.2)	
Mean pack-year of smoker (range)	36.8 (2.5-130)	34.8 (0.5-120)	0.54
Synchronicity, n (%)			0.28
Synchronous	103 (66.0)	198 (71.5)	
Metachronous	53 (34.0)	79 (28.5)	
Median time interval for metachronous case, m (range)	24.7 (0-129.8)	23.6 (0-142.3)	0.49
Distribution of tumors, n (%)			**0.004**
Ipsilateral (same lobe)	67 (43.0)	86 (31.1)	
Ipsilateral (different lobe)	57 (36.5)	95 (34.3)	
Contralateral	32 (20.5)	96 (34.7)	
Lung cancer stage group, 8th edition	(available N=147)	(available N=271)	**<0.001**
Stage 0		3(1.2)	
Stage I		203 (73.2)	
Stage II	40 (27.2)	37 (14.6)	
Stage III	76 (51.7)	27 (10.6)	
Stage IV	31 (21.1)	1(0.4)	
Median tumor size, cm (range)	3.05 (0.1-15)	2.3 (0.1-9.2)	**<0.001**
The average ISALC grade	2.27	1.70	**<0.001**
Nodal status			**0.003**
N0 or Nx	111 (71.2)	231 (83.4)	
N1/2/3	45 (28.8)	46 (16.6)	

MPLC: multiple primary lung cancer, IPM: intrapulmonary metastasis.

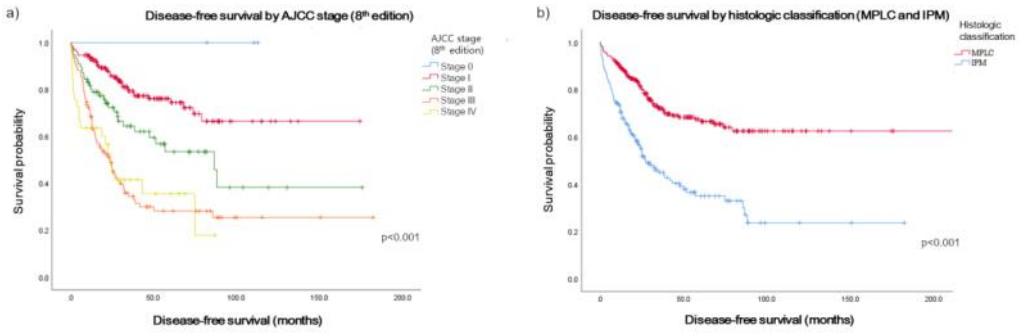


Figure 3. Kaplan–Meier survival curves for disease-free survival of non-small cell lung cancer patients with multiple tumors resected at Yonsei University Severance Hospital (2006-2020). a) based on AJCC stage ($p<0.001$) and b) histologic classification ($p<0.001$).

In a cohort of 433 non-small cell lung cancer (NSCLC) patients with multiple tumors resected at Yonsei University Severance Hospital (Seoul, Korea) between 2006 and 2020, we identified several statistically significant differences between MPLC and IPM groups when classified based on histologic criteria. (Table 3)

In the IPM group (n=156), there was a higher proportion of males (104, 66.7%) compared to the MPLC group (n=277) with 143 males (51.6%) ($p<0.001$). Additionally, the IPM group had a lower percentage of non-smokers (54, 34.6%) compared to the MPLC group (153, 55.2%) ($p<0.001$). Patients in the IPM group underwent resection for their first tumor at a slightly younger median age (64 years) compared to those in the MPLC group (65 years) ($p=0.027$).

In terms of tumor distribution, the IPM group showed a decreasing proportion from intralobal (43.0%) to different but ipsilateral lobe (36.5%) and to contralateral lobe (20.5%). In contrast, the MPLC group exhibited a relatively even distribution across the locations with 31.1% in intralobal, 34.3% in different but ipsilateral lobe, and 34.7% in contralateral lobe ($p=0.004$).

Regarding other prognostic factors, the IPM group consistently demonstrated poorer

prognosis compared to the MPLC group: IPM had a larger median tumor size (3.05cm vs 2.3cm) ($p<0.001$), a higher average ISALC grade for non-mucinous adenocarcinoma cases (2.27 vs 1.70) ($p<0.001$), and a greater proportion of nodal metastasis (28.8% vs 16.6%) ($p=0.003$). In the final AJCC stage classification, the proportion of patients in advanced stages (stage III-IV) was 72.8% for IPM and 11% for MPLC ($p<0.001$). Most cases showed poor disease-free survival as the pathologic stage increased ($p<0.001$) (Fig. 3a), and as a result, IPM cases demonstrated shorter disease-free survival compared to MPLC cases ($p<0.001$). (Fig. 3b)

Among the study group, 62 patients had available information on the driver mutation status (EGFR, ALK and ROS1) for all of their tumors; 2 tumors in 58 patients and 3 tumors in 4 patients, totaling 128 tumors. The most common mutation was the EGFR mutation (80, 62.5%), with the L858R missense mutation (49, 38.3%) being the most frequently detected. This was followed by the ROS1 (3, 2.4%) and ALK (1, 0.8%) rearrangements. Notably, in one case of ROS1, an EGFR Exon 19 deletion mutation was detected concomitantly. (Supplementary table 3)

4. Comparison to histologic classification and MeTel algorithm in in-house data

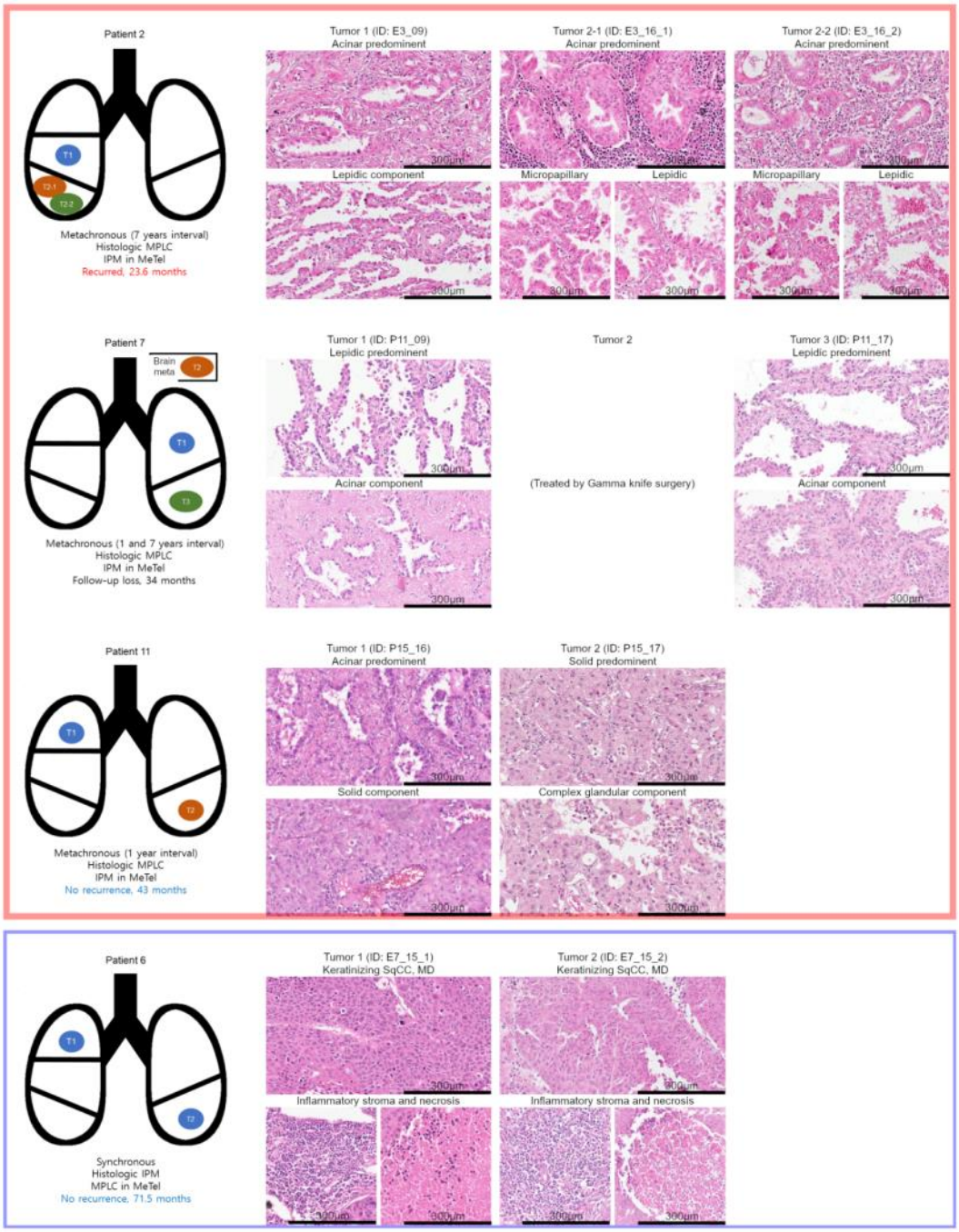


Figure 4. Discordant cases between histologic predictions and MeTel analysis. Tumors

with distinct histologic histology are represented by different colors, while those with similar histologic characteristics share the same color. Each tumor is sequentially numbered according to the order of occurrence. The microscopic slide on the top displays the major histologic patterns of each tumor, and the bottom displays other secondary histological components. MPLC: multiple primary lung cancer, IPM: intrapulmonary metastasis, SqCC: squamous cell carcinoma, MD: moderately differentiated.

In comparison of histologic classification and MeTel algorithm, 33.3% (2 out of 6) of the patients in both the WES and TSO500 datasets showed disagreement in the prediction of MPLC or IPM. (Fig. 4 and Supplementary table 7) For patient 9, exon 19 deletions in both tumors confirmed via TSO500 panel sequencing, concluding that the previous EGFR sequencing of the first tumor (EGFR wild) resulted in a false negative. This resulted in a consistent IPM diagnosis from both the histologic classification and MeTel algorithm. In the WES dataset, Patient 2 was classified as MPLC in the histologic classification but as IPM in the MeTel algorithm and another patient (Patient 6) had the opposite classification. In the TSO500 dataset, both Patients 7 and 11 were classified as MPLC in the histologic classification but as IPM in the MeTel algorithm. The confidence level of the MeTel algorithm for all TSO500 and WES dataset samples was ‘confident'.

In Patient 2, three tumors were resected: one adenocarcinoma in the right middle lobe in 2009 and two synchronous adenocarcinomas in the right lower lobe in 2016. During histologic evaluation, all tumors were found to have a lepidic component (10-40%), which was interpreted as a non-invasive precursor lesion and therefore classified as MPLC. However, the MeTel algorithm classified all tumors as IPM. Upon re-review, the two tumors in 2016 were found to be adenocarcinomas with high-grade histologic components (10-15% micropapillary), while the first tumor in 2009 was an adenocarcinoma with only acinar (60%) and lepidic (40%) components. The patient was found to have subdiaphragmatic metastasis on an abdominal CT two years after the last resection of the tumors.

Patient 6 was diagnosed with two synchronous squamous cell carcinomas located in

the left lower lobe and right upper lobe in 2015. Microscopic examination revealed that both tumors displayed similar histologic features, including keratinizing type, moderate differentiation, necrosis (10-30%), and an inflammatory stroma, but no evidence of a precursor lesion such as squamous dysplasia. Based on these findings, the tumors were classified as IPM in the histologic evaluation. A clonality test using WES data showed distinct copy number variations in both tumors, suggesting that their true nature was MPLC. The patient has been under observation for 71.5 months without recurrence.

In the case of Patient 7, two cases of adenocarcinoma were detected with an interval of 8 years. The first tumor was identified in the left upper lobe in 2009 and exhibited a lepidic predominant (70%) pattern. After one year, a brain metastasis was found, which was treated with gamma knife surgery and subsequently Navelbine chemotherapy. In 2017, a second tumor was discovered in the left lower lobe, which also had a lepidic predominant (90%) pattern, leading to MPLC classification by histologic criteria. However, the presence of systemic metastasis raises the possibility that the tumors may be IPM. MeTel classified the tumor pair as IPM, and further NGS analysis also showed similar copy number variations and eight shared mutations between the tumors, favoring IPM. The patient received palliative care at another hospital without additional chemotherapy, and follow-up was lost 34 months after the last resection.

In Patient 11, a lung adenocarcinoma first occurred in the right upper lobe in 2016 and another case of lung adenocarcinoma was found in the left lower lobe a year later in 2017. The first tumor showed an acinar predominant (60%) pattern with minor solid (30%) and micropapillary (10%) components. The second tumor showed a much higher solid component (80%), resulting in a diagnosis of MPLC due to the difference in predominant histologic type. MeTel classified the tumor pair as IPM, and further NGS analysis revealed similar chromosomal patterns, with copy number gains on chromosomes 5 and 7 and a shared rare genomic variant (AKT3 c.*5422T>A),

supporting MeTel's result. The patient has not exhibited any recurrence for over 43 months during the current follow-up period.

5. The Ethnic-Specific Mode of MeTel

Table 4. Accuracy of MeTel with race information.

Dataset (Country of the institution where the sample was collected)	Accuracy with race information (%)			
	Asian	Black	White	Total population
22 (France)	94.50	95.41	95.41	95.41
189 (France)	100	100	100	100
468 (America)	100	100	100	100
520 (China)	100	100	100	100
523[1] (South Korea)	100	100	100	100
605 (China)	98.04	92.16	92.16	96.08
WES-1 (China)	100	100	100	100
WES-2[1] (South Korea)	100	100	100	100
Total	97.61	96.93	96.93	97.61

The country of the institution from which the samples were collected and the accuracy of each dataset.
[1] in-house dataset from Severance Hospital.

To further enhance the sophistication of MeTel, we propose incorporating ethnic-specific differences in mutation frequency into the calculation of the likelihood formula. Due to significant variations in mutation frequency among different ethnic populations, considering these differences is expected to improve the accuracy of MeTel's classification.

To evaluate the impact of the ethnic-specific mode of MeTel, we applied the value p that reflects such differences and assessed the accuracy for each dataset. While ethnic information for each sample was not available in most datasets, since the nationality of the institution from which each sample was collected was known, we could infer the ethnic composition of each dataset based on the nationality of the collecting institution (Table 4).

Our analysis revealed interesting findings. When considering the Asian population frequency, we observed a decrease in accuracy for the 22-panel dataset obtained from institutions in France, but an increase for the 605-panel dataset from Chinese institutions. Conversely, when applying non-Asian population frequencies, such as Black and White, we observed a decrease in accuracy for the 605-panel dataset. Notably, specific mutations, such as KRAS (p.G12C) in the 22-panel dataset and EGFR (p.L858R) in the 605-panel dataset, showed significant differences in frequency based on race, resulting in different classification outcomes compared to the original MeTel results (Supplementary Table 8 and Supplementary Fig. 2).

These findings suggest that incorporating ethnicity information of patients into MeTel's process could enhance its sophistication and accuracy. Currently, in this study, MeTel relies on the frequency of mutations in the entire population, but when using MeTel, it is recommended to utilize the ethnic-specific mode to better reflect the genetic diversity across different populations.

IV. DISCUSSION

1. MeTel Algorithm

In this study, we introduced MeTel, a novel sequencing-based classification algorithm for identifying IPM and MPLC in patients with multiple lung cancer. MeTel overcomes the limitations of previous genomic methods by estimating the probabilities of IPM and MPLC based on the somatic mutation profile of tumors, rather than relying solely on the number of shared mutations.

The distinct error trends of each algorithm and accuracy rate by panel size in the classification result (Fig.1b and 1d) shows the limitation of previous mutation count-based algorithms and outstanding performance of probability ratio-based algorithm, MeTel, especially in large-sized gene panel. With the count-based algorithm, as the panel size decreases, there's a reduced opportunity to detect shared mutations. This reduction manifests as an increasing rate of IPM errors. Conversely, in larger panels, the incidence of coinciding MPLC mutations rises, triggering a surge in MPLC errors. As the gene panel size expands to whole-exome size, the accuracy of count-based algorithm dramatically decreases, causing errors in most cases. It's crucial to note that these two error types have opposing influences: this interplay means that merely adjusting the cutoff of mutation count won't optimize overall accuracy. Consequently, when charting the accuracy rate against panel size, we witness a reverse U-shaped trend of previous mutation count-based algorithms.

MeTel employs a Bayesian probability model to quantitatively calculate and compare the likelihood of coincidental mutation matching in MPLC and the potential for transition in IPM for each detected mutation. MeTel's analytical strength intensifies as the number of mutations detected within the panel increases, and it is free from the MPLC errors in large panels caused by count-based algorithm. This is evident in the near-perfect performance of MeTel when compared to the significantly lower accuracy of the count-based algorithm in the whole exome panel cohort.

MeTel's performance consistently outperformed other algorithms, even in smaller-sized panels like the 22 genes. The few errors that were observed with MeTel indicated a distinct algorithmic principle compared to other count-based algorithms. While the primary error from count-based algorithms in small panels arises from insufficient detection of shared mutations in IPM, causing IPM error, all of MeTel's error instances occurred in the study where only a few mutations were detected that shared single driver gene mutation. Consequently, there was a mistaken interpretation of coincidentally matching driver gene mutations in MPLC as evidence for IPM, leading to MPLC errors.

The impact of ethnicity also plays a role in these errors. Since the mutation incidence was calculated based on an overall population without considering ethnicity, the probability of coincidence for driver gene mutations that frequently appear within specific ethnic cohorts was underestimated, leading to incorrect classifications as IPM. This can be corroborated when applying the updated incidences specific to each ethnicity for KRAS and EGFR mutations, as seen in cases where the errors were corrected. (Supplementary Table 8 and Supplementary fig. 2)

As mentioned previously, such errors occurred when the number of detected mutations was insufficient to provide ample evidence for classification. We were able to categorize potential risk groups by distinguishing confidence levels in a tiered manner. When cases were divided into three levels based on the MAX (p_i, p_m) value (Likely, Probable, and Confident), all the error cases in this study fell into the group with the lowest confidence, the 'Likely' group. Enhancing the reliability of these cases by optionally applying classic histologic criteria resulted in an accuracy level comparable to that of the other confidence groups. Similarly, among the count-based algorithms, the MSK model employs a histology-molecular combined algorithm. When evaluating tumor pairs, it first applies histologic criteria to determine distinct MPLC. Compared to the other two algorithms, the MSK model demonstrated higher accuracy in the whole exome panel.

We applied MeTel to 12 patients at our institution whose classifications were challenging based solely on pathological findings. Among them, four showed discordant result with MeTel classification. Upon conducting additional NGS analysis, such as copy number variation, the result favored MeTel's categorization. This demonstrates MeTel's potential for real-world clinical trial and its excellence as a supplementary tool complementing traditional histological classifications.

There are several limitations to consider before implementing MeTel in actual clinical settings. Firstly, in Korea, only a few driver genes (EGFR, ALK, ROS1) targeting lung tumors are currently sequenced. Due to medical insurance policies, only one of the tumors is allowed NGS-scale molecular test. As such, to fully utilize MeTel, additional molecular testing is inevitable. Secondly, in small tissue samples, like biopsies, tumor purity can greatly vary due to the presence of normal contaminants such as inflammatory cells. Compared to resection specimens, there’s a need to consider the tumor purity as a significant variable.

Fortunately, the NGS test for NSCLC patients with 23-gene panels (Oncomine Dx target test)[32] has been approved by medical insurance in Korea since december 2022. As MeTel has demonstrated high performance in this scale of panel, we anticipate its increased applicability in actual clinical environments in the future.

2. Pathologic review of discordant cases between histologic classification and MeTel algorithm

The time interval serves as one of the key indicators in distinguishing between MPLC and IPM in cases of multiple lung cancer. The majority of NSCLC cases recur within two years of onset, hence the two-year cutoff in the Martini criteria[7] and the extended four-year cutoff in the ACCP [8] were established based on such clinical incidence data. The Cochin method also utilizes a five-year cutoff for classifying MPLC [13]. In our cohort of multiple lung cancer patients, cases classified as IPM based on histologic

findings were assembled, we found that approximately 94% of them had a time interval of less than five years (Supplementary table 9).

Despite the convenience of analysis, the time interval cutoff, based on incidence-based statistics, has inherent limitations in terms of fundamental reliability. In our study, all cases selected with a time interval exceeding five years (patients 1-5, 7-8) were classified as IPM in the in-depth classification by MeTel. Given the clinical significance that the distinction between MPLC and IPM holds for patients, it is necessary to be cautious about using time intervals as classification criteria.

In the cases of patients 2 and 7, both patients were initially classified as MPLC based on the presence of a lepidic (in situ) component in their respective tumors, not only based on time interval. Irrespective of the type and organ of origin, such precancerous lesions have consistently been employed as strong evidence for primary cancer, rather than metastasis.[33-35] Similarly, in conventional histologic criteria for lung cancer, the presence of an in situ component plays a pivotal role in distinguishing primary cancer from metastasis.[23, 36] However, these cases were reclassified as IPM using MeTel.

In lung adenocarcinoma, the in situ component is characterized by neoplastic cells confined to the pre-existing alveolar wall, absent any structural destruction [23, 36]. This growth pattern is now termed 'lepidic', and non-mucinous lung adenocarcinoma with a purely or predominantly lepidic pattern is classified as adenocarcinoma in situ (AIS), minimally invasive adenocarcinoma (MIA), and lepidic adenocarcinoma, depending on the presence and size of invasive foci[23]. However, certain invasive adenocarcinomas can demonstrate outgrowth along the alveolar wall, mimicking precancerous lesions. Moore *et al*. suggested several features that represent peripheral outgrowth of the invasive component rather than low-grade precursor: (1) clear nuclear difference between invasive and in situ component; (2) architectural asymmetry of invasive component; and (3) the absence of lepidic “penumbra”, composing a uniformly widened thickness of the in situ component.[37] In a recent study conducted by the

IASLC pathology committee, several pathologists derived potentially useful features for distinguishing between in situ components and invasive patterns through the Delphi approach.[38] According to their findings, the invasive pattern includes two or more of the following three major histologic criteria: (1) "Extensive Epithelial Proliferation" (EEP), (2) desmoplasia, and (3) altered alveolar structure. EEP is defined as epithelial cells, consisting of more than two cell layers, growing along the alveolar wall accompanied by cytologic atypia. If these major criteria are not met, the presence of high-grade cytologic atypia, cytologic transition between the invasive and in situ components, or the presence of macrophages within a collapsed space are considered as useful features to determine "considerable" invasive pattern. On the other hand, monolayered lepidic component with compressed but regular parallel or streaming pattern was considered as iatrogenic collapse, favoring non-invasive pattern.

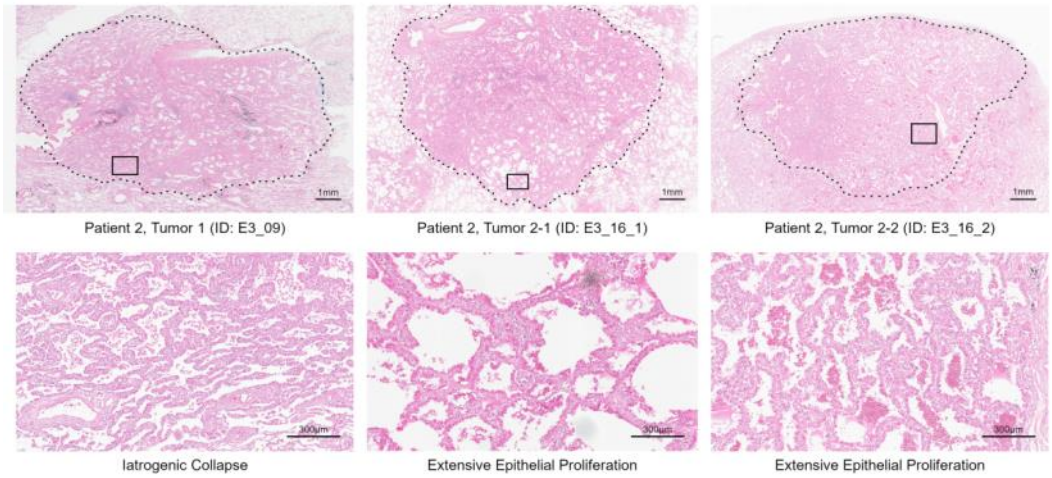


Figure 5. Lepidic components of the early and later tumors from Patient 2. The top image provides an overview of the entire tumor, with the tumor's outer boundary delineated by dotted lines. The bottom image shows a higher magnification of the lepidic component area marked by a rectangle in the overview image. While Tumor 1 displayed a typical iatrogenic collapse pattern, Tumor 2-1 and Tumor 2-2 exhibited features approaching extensive epithelial proliferation.

For the cases of Patient 2, there were differences in the morphology of the lepidic component between the tumor from 2009 and the two tumors from 2016. (Fig. 5) The

early tumor from 2009 displayed a typical iatrogenic collapse pattern, with the lepidic component showing low-grade atypia being compressed in parallel direction, and occasional alveolar macrophages observed in the lumen. However, the lepidic component of the later tumors from 2016 showed irregular fibrosis of the alveolar wall and stratification of more than 2 cells, considered as EEP. Moreover, these components partly seemed to transit into micropapillary components, suggesting the possibility that the lepidic component of the later tumors might be an outgrowth of an invasive pattern.

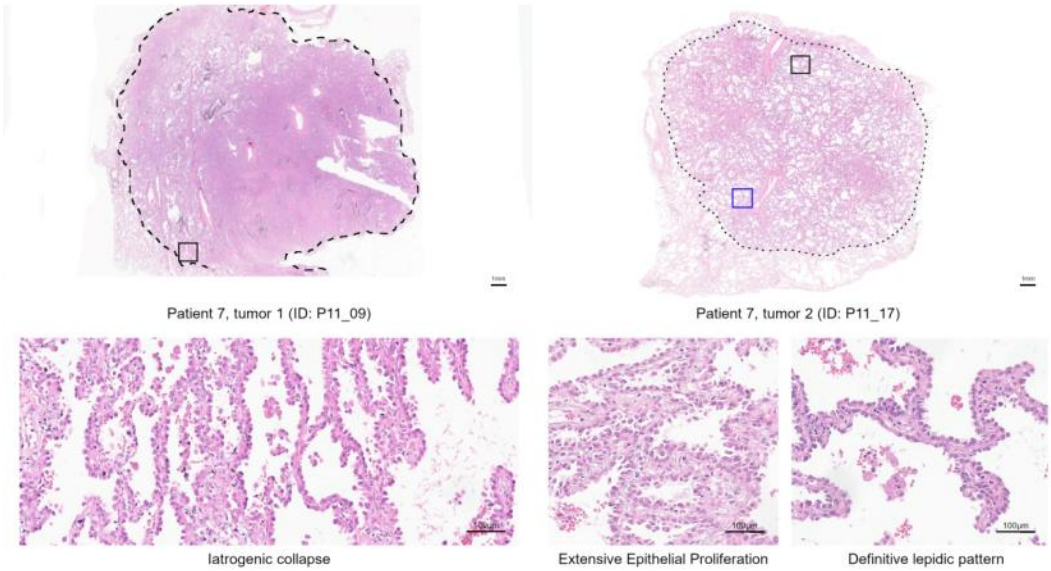


Figure 6. Lepidic component of the early and later tumor from Patient 7. The top image provides an overview of the entire tumor, with the tumor's outer boundary and the invasive (acinar) pattern indicated by dotted lines. The bottom image shows a higher magnification of the extensive epithelial proliferation (EEP) and non-EEP areas of the tumor's lepidic component, each marked by blue and black rectangles, respectively, in the overview image. Histologically, early tumor shows iatrogenic collapse pattern, but later tumor exhibits a mix of EEP and definitive lepidic pattern, leading it to be an equivocal case.

However, in the case of Patient 7, the early tumor shows iatrogenic collapse pattern but the later tumor is histologically mixture of EEP and definitive lepidic pattern, leading it to be an equivocal case. (Fig. 6) Previous study already showed that the lepidic component does not always guarantee MPLC through the discordance with molecular testing, but the result was only restricted to lepidic non-predominant cases.[14] On the other hand, the brain metastasis history could provide a clinical clue suggesting IPM. Considering the probability that the mutations confirmed in TSO500 sequencing coincidentally match, it would be more rational to favor the possibility of IPM.

The ability to confirm the metastatic nature of squamous cell carcinoma in the case of Patient 6 provides an encouraging perspective. Compared with lung adenocarcinoma, pulmonary squamous cell carcinoma displays a relatively lower frequency of driver gene alterations such as EGFR mutations or ALK rearrangements. [22] Beyond the rarity of these specific gene alterations, the histological diversity of lung squamous cell carcinoma is also less complex than that of lung adenocarcinoma, dividing into only three categories: keratinizing, non-keratinizing, and basaloid. [23] When pulmonary squamous cell carcinoma presented as multiple intrapulmonary tumors, such lack of the significant driver gene and morphologic heterogeneity make it difficult to predict clonal relationship of the tumors.[39]

In histological assessment for multiple pulmonary squamous cell carcinoma, detailed morphologic characteristics beyond the WHO classification was suggested to distinguish MPLC and IPM: degree of keratinization, necrosis, desmoplasia and inflammation and rare cytologic features such as clear cell, papillary, basaloid or sarcomatoid patterns.[9] However, in our study, the squamous cell carcinoma pair demonstrated indistinguishable cytologic, structural, and stromal backgrounds, making it impossible to differentiate MPLC. Using WES analysis to examine the copy number variation data, we confirmed that these tumors are independently occurred. This finding was consistent with the patient's outcome, who has remained recurrence-free more than 5 years.

In Patient 11, the discrepancy in predominant histologic components between two tumors led to a classification as MPLC. However, detailed histological review revealed this was due to an increased high-grade (solid) component in the secondary tumor that appeared a year later. The International Association for the Study of Lung Cancer (IASLC) categorized these high-grade patterns - solid, micropapillary, and complex glandular - in 2020, and established their correlation with poorer prognosis in non-mucinous adenocarcinoma.[40, 41] Previous reports also documented increased proportions of these high-grade components in metastatic cancers.[14, 42] Through molecular analysis using the MeTel model, we were able to rectify the histologic discrepancy by confirming that the two tumors shared unique genomic variants. Consequently, variations in predominant histologic patterns should not be hastily interpreted as MPLC, but rather carefully assessed for potential progression during metastasis.

3. Potential candidates for the expanded application of MeTel

In this study, the performance of MeTel was assessed using a dataset predominantly composed of non-mucinous adenocarcinoma cases. However, since it operates without being limited to specific cancer types or driver genes, there is potential to expand its application.

Pulmonary mucinous adenocarcinoma exhibits a histology similar to pancreatobiliary adenocarcinoma. Both have a high frequency of KRAS mutations, making differentiation between metastasis or double primary challenging.[43, 44] Although TTF-1 marker is known to be relatively specific to pulmonary mucinous adenocarcinoma[45], its expression frequency isn't high, underscoring the need for a more accurate differential analysis.

Squamous cell carcinoma does not present as diverse morphologic variations as

adenocarcinoma, making the application of classic histologic criteria challenging.[23] However, lung squamous cell carcinoma generally has a higher tumor mutation burden (TMB) than lung adenocarcinoma[46], suggesting a greater utility for molecular-based classification like MeTel.

Head and neck squamous cell carcinoma (HNSCC) often metastasizes to the lungs[47], but double primary lung squamous cell carcinoma is also frequently observed, appearing in about 5-19% of cases[48]. While considering surgical resection for double primary lung squamous cell carcinoma, the direction of treatment may shift towards palliative therapy for metastatic HNSCC.[49, 50] At present, there is a lack of clear criteria or biomarkers to differentiate the two[50], but with a sufficient cohort, the application of MeTel can be anticipated.

V. CONCLUSION

Identifying the clonality of multiple lung cancers is crucial for establishing an accurate T stage and treatment plan. MeTel offers consistent and accurate analysis, surpassing previously reported algorithms. Furthermore, the integration of MeTel can help to refine traditional criteria based on morphological analysis and result in more accurate prognostic predictions.

REFERENCES

1. Karacz CM, Yan J, Zhu H, Gerber DE. Timing, Sites, and Correlates of Lung Cancer Recurrence. Clin Lung Cancer. 2020;21(2):127-35.e3. Epub 2020/01/15. doi: 10.1016/j.cllc.2019.12.001. PubMed PMID: 31932216; PubMed Central PMCID: PMC7061059.
2. Sasaki H, Suzuki A, Tatematsu T, Shitara M, Hikosaka Y, Okuda K, et al. Prognosis of recurrent non-small cell lung cancer following complete resection. Oncol Lett. 2014;7(4):1300-4. Epub 2014/06/20. doi: 10.3892/ol.2014.1861. PubMed PMID: 24944713; PubMed Central PMCID: PMC3961426.
3. Ichinose Y, Yano T, Asoh H, Yokoyama H, Yoshino I, Katsuda Y. Prognostic factors obtained by a pathologic examination in completely resected non-small-cell lung cancer. An analysis in each pathologic stage. J Thorac Cardiovasc Surg. 1995;110(3):601-5. doi: 10.1016/s0022-5223(95)70090-0. PubMed PMID: 7564425.
4. Rea F, Zuin A, Callegaro D, Bortolotti L, Guanella G, Sartori F. Surgical results for multiple primary lung cancers. Eur J Cardiothorac Surg. 2001;20(3):489-95. doi: 10.1016/s1010-7940(01)00858-2. PubMed PMID: 11509268.
5. Amin MB, Edge SB, Greene FL, Byrd DR, Brookland RK, Washington MK, et al. AJCC Cancer Staging Manual: Springer International Publishing; 2018.
6. National Comprehensive Cancer Network. NCCN Clinical Practice Guidelines in Oncology (NCCN Guidelines®): Non-small cell lung cancer (version 4.2022) 2022 [cited 2022 02/09]. Available from: https://www.nccn.org/professionals/physician_gls/pdf/nscl.pdf.
7. Martini N, Melamed MR. Multiple primary lung cancers. J Thorac Cardiovasc Surg. 1975;70(4):606-12. Epub 1975/10/01. PubMed PMID: 170482.
8. Shen KR, Meyers BF, Larner JM, Jones DR. Special treatment issues in lung cancer: ACCP evidence-based clinical practice guidelines (2nd edition). Chest. 2007;132(3 Suppl):290s-305s. Epub 2007/10/06. doi: 10.1378/chest.07-1382. PubMed PMID: 17873175.
9. Girard N, Deshpande C, Lau C, Finley D, Rusch V, Pao W, et al. Comprehensive histologic assessment helps to differentiate multiple lung primary nonsmall cell carcinomas from metastases. Am J Surg Pathol. 2009;33(12):1752-64. Epub 2009/09/24. doi: 10.1097/PAS.0b013e3181b8cf03. PubMed PMID: 19773638; PubMed Central PMCID: PMC5661977.
10. Nicholson AG, Torkko K, Viola P, Duhig E, Geisinger K, Borczuk AC, et al. Interobserver Variation among Pathologists and Refinement of Criteria in Distinguishing Separate Primary Tumors from Intrapulmonary Metastases in Lung. J Thorac Oncol. 2018;13(2):205-17. Epub 2017/11/12. doi: 10.1016/j.jtho.2017.10.019. PubMed PMID: 29127023; PubMed Central PMCID: PMC6276791.
11. Chang YL, Wu CT, Lin SC, Hsiao CF, Jou YS, Lee YC. Clonality and prognostic implications of p53 and epidermal growth factor receptor somatic aberrations in multiple primary lung cancers. Clin Cancer Res. 2007;13(1):52-8. Epub 2007/01/04. doi: 10.1158/1078-0432.Ccr-06-1743. PubMed PMID: 17200338.

12. Jiang L, He J, Shi X, Shen J, Liang W, Yang C, et al. Prognosis of synchronous and metachronous multiple primary lung cancers: systematic review and meta-analysis. Lung Cancer. 2015;87(3):303-10. Epub 2015/01/27. doi: 10.1016/j.lungcan.2014.12.013. PubMed PMID: 25617985.
13. Mansuet-Lupo A, Barritault M, Alifano M, Janet-Vendroux A, Zarmaev M, Biton J, et al. Proposal for a Combined Histomolecular Algorithm to Distinguish Multiple Primary Adenocarcinomas from Intrapulmonary Metastasis in Patients with Multiple Lung Tumors. J Thorac Oncol. 2019;14(5):844-56. Epub 2019/02/06. doi: 10.1016/j.jtho.2019.01.017. PubMed PMID: 30721797.
14. Chang JC, Alex D, Bott M, Tan KS, Seshan V, Golden A, et al. Comprehensive Next-Generation Sequencing Unambiguously Distinguishes Separate Primary Lung Carcinomas From Intrapulmonary Metastases: Comparison with Standard Histopathologic Approach. Clin Cancer Res. 2019;25(23):7113-25. Epub 2019/09/01. doi: 10.1158/1078-0432.Ccr-19-1700. PubMed PMID: 31471310; PubMed Central PMCID: PMC7713586.
15. Wang X, Gong Y, Yao J, Chen Y, Li Y, Zeng Z, et al. Establishment of Criteria for Molecular Differential Diagnosis of MPLC and IPM. Front Oncol. 2020;10:614430. Epub 20210121. doi: 10.3389/fonc.2020.614430. PubMed PMID: 33552986; PubMed Central PMCID: PMC7860975.
16. Suh YJ, Lee HJ, Sung P, Yoen H, Kim S, Han S, et al. A Novel Algorithm to Differentiate Between Multiple Primary Lung Cancers and Intrapulmonary Metastasis in Multiple Lung Cancers With Multiple Pulmonary Sites of Involvement. J Thorac Oncol. 2020;15(2):203-15. Epub 20191018. doi: 10.1016/j.jtho.2019.09.221. PubMed PMID: 31634666.
17. Shao J, Wang C, Li J, Song L, Li L, Tian P, et al. A comprehensive algorithm to distinguish between MPLC and IPM in multiple lung tumors patients. Ann Transl Med. 2020;8(18):1137. doi: 10.21037/atm-20-5505. PubMed PMID: 33240986; PubMed Central PMCID: PMC7576050.
18. Duan J, Ge M, Peng J, Zhang Y, Yang L, Wang T, et al. Application of large-scale targeted sequencing to distinguish multiple lung primary tumors from intrapulmonary metastases. Sci Rep. 2020;10(1):18840. Epub 20201102. doi: 10.1038/s41598-020-75935-4. PubMed PMID: 33139840; PubMed Central PMCID: PMC7606457.
19. Liu Y, Zhang J, Li L, Yin G, Zhang J, Zheng S, et al. Genomic heterogeneity of multiple synchronous lung cancer. Nat Commun. 2016;7:13200. Epub 20161021. doi: 10.1038/ncomms13200. PubMed PMID: 27767028; PubMed Central PMCID: PMC5078731.
20. Vignot S, Frampton GM, Soria J-C, Yelensky R, Commo F, Brambilla C, et al. Next-Generation Sequencing Reveals High Concordance of Recurrent Somatic Alterations Between Primary Tumor and Metastases From Patients With Non–Small-Cell Lung Cancer. Journal of Clinical Oncology. 2013;31(17):2167-72. doi: 10.1200/JCO.2012.47.7737.
21. Tian H, Wang Y, Yang Z, Chen P, Xu J, Tian Y, et al. Genetic trajectory and clonal evolution of multiple primary lung cancer with lymph node metastasis. Cancer Gene Ther. 2023;30(3):507-20. Epub 20230119. doi: 10.1038/s41417-022-00572-0. PubMed PMID:

36653483; PubMed Central PMCID: PMC10014582.
22. Rekhtman N, Paik PK, Arcila ME, Tafe LJ, Oxnard GR, Moreira AL, et al. Clarifying the spectrum of driver oncogene mutations in biomarker-verified squamous carcinoma of lung: lack of EGFR/KRAS and presence of PIK3CA/AKT1 mutations. Clin Cancer Res. 2012;18(4):1167-76. Epub 2012/01/10. doi: 10.1158/1078-0432.Ccr-11-2109. PubMed PMID: 22228640; PubMed Central PMCID: PMC3487403.
23. Board WCoTE. WHO classification of tumours: Thoracic tumours. Lyon: World Health Organization (WHO); 2021. 90-1 p.
24. Zhao C, Jiang T, Ju JH, Zhang S, Tao J, Fu Y, et al. TruSight Oncology 500: Enabling Comprehensive Genomic Profiling and Biomarker Reporting with Targeted Sequencing. bioRxiv. 2020:2020.10.21.349100. doi: 10.1101/2020.10.21.349100.
25. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nature Biotechnology. 2011;29(1):24-6. doi: 10.1038/nbt.1754. PubMed PMID: WOS:000286048900013.
26. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv: Genomics. 2013.
27. Van der Auwera GA OCB. Genomics in the Cloud: Using Docker, GATK, and WDL in Terra. O'Reilly Media. 2020.
28. Diossy M, Sztupinszki Z, Krzystanek M, Borcsok J, Eklund AC, Csabai I, et al. Strand Orientation Bias Detector to determine the probability of FFPE sequencing artifacts. Briefings in Bioinformatics. 2021;22(6). doi: ARTN bbab186
10.1093/bib/bbab186. PubMed PMID: WOS:000733325700056.
29. Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E, Kallberg M, et al. Strelka2: fast and accurate calling of germline and somatic variants. Nature Methods. 2018;15(8):591-+. doi: 10.1038/s41592-018-0051-x. PubMed PMID: WOS:000440334000017.
30. Koboldt DC. Best practices for variant calling in clinical sequencing. Genome Med. 2020;12(1):91. Epub 20201026. doi: 10.1186/s13073-020-00791-w. PubMed PMID: 33106175; PubMed Central PMCID: PMC7586657.
31. McHugh ML. Interrater reliability: the kappa statistic. Biochemia Medica. 2012;22(3):276-82. doi: DOI 10.11613/bm.2012.031. PubMed PMID: WOS:000309525200005.
32. Administration USFaD. Oncomine™ Dx Target Test Part I: Sample Preparation and Quantification User Guide. Revision C.0 ed2017.
33. Lee AH. The histological diagnosis of metastases to the breast from extramammary malignancies. J Clin Pathol. 2007;60(12):1333-41. Epub 2007/11/29. doi: 10.1136/jcp.2006.046078. PubMed PMID: 18042689; PubMed Central PMCID: PMC2095576.
34. Saida T, Tanaka YO, Matsumoto K, Satoh T, Yoshikawa H, Minami M. Revised FIGO staging system for cancer of the ovary, fallopian tube, and peritoneum: important implications for radiologists. Japanese Journal of Radiology. 2016;34(2):117-24. doi: 10.1007/s11604-015-0513-3.

35. Jiang K, Al-Diffalha S, Centeno BA. Primary Liver Cancers—Part 1:Histopathology, Differential Diagnoses, and Risk Stratification. Cancer Control. 2018;25(1):1073274817744625. doi: 10.1177/1073274817744625. PubMed PMID: 29350068.
36. Borczuk AC. Assessment of invasion in lung adenocarcinoma classification, including adenocarcinoma in situ and minimally invasive adenocarcinoma. Modern Pathology. 2012;25(1):S1-S10. doi: 10.1038/modpathol.2011.151.
37. Moore DA, Sereno M, Das M, Baena Acevedo JD, Sinnadurai S, Smith C, et al. In situ growth in early lung adenocarcinoma may represent precursor growth or invasive clone outgrowth—a clinically relevant distinction. Modern Pathology. 2019;32(8):1095-105. doi: 10.1038/s41379-019-0257-1.
38. Thunnissen E, Beasley MB, Borczuk A, Dacic S, Kerr KM, Lissenberg-Witte B, et al. Defining Morphologic Features of Invasion in Pulmonary Nonmucinous Adenocarcinoma With Lepidic Growth: A Proposal by the International Association for the Study of Lung Cancer Pathology Committee. Journal of Thoracic Oncology. 2023;18(4):447-62.
39. Schneider F, Derrick V, Davison JM, Strollo D, Incharoen P, Dacic S. Morphological and molecular approach to synchronous non-small cell lung carcinomas: impact on staging. Modern Pathology. 2016;29(7):735-42. doi: 10.1038/modpathol.2016.66.
40. Travis WD, Brambilla E, Noguchi M, Nicholson AG, Geisinger KR, Yatabe Y, et al. International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society International Multidisciplinary Classification of Lung Adenocarcinoma. Journal of Thoracic Oncology. 2011;6(2):244-85. doi: https://doi.org/10.1097/JTO.0b013e318206a221.
41. Woo W, Cha Y-J, Kim BJ, Moon DH, Lee S. Validation Study of New IASLC Histology Grading System in Stage I Non-Mucinous Adenocarcinoma Comparing With Minimally Invasive Adenocarcinoma. Clinical Lung Cancer. 2022. doi: https://doi.org/10.1016/j.cllc.2022.06.004.
42. Aokage K, Ishii G, Yoshida J, Hishida T, Nishimura M, Nagai K, et al. Histological progression of small intrapulmonary metastatic tumor from primary lung adenocarcinoma. Pathology International. 2010;60(12):765-73. doi: https://doi.org/10.1111/j.1440-1827.2010.02596.x.
43. Krasinskas AM, Chiosea SI, Pal T, Dacic S. KRAS mutational analysis and immunohistochemical studies can help distinguish pancreatic metastases from primary lung adenocarcinomas. Mod Pathol. 2014;27(2):262-70. Epub 20130726. doi: 10.1038/modpathol.2013.146. PubMed PMID: 23887294; PubMed Central PMCID: PMC4091042.
44. Cha YJ, Shim HS. Biology of invasive mucinous adenocarcinoma of the lung. Transl Lung Cancer Res. 2017;6(5):508-12. doi: 10.21037/tlcr.2017.06.10. PubMed PMID: 29114467; PubMed Central PMCID: PMC5653530.
45. Hwang DH, Sholl LM, Rojas-Rudilla V, Hall DL, Shivdasani P, Garcia EP, et al.

KRAS and NKX2-1 Mutations in Invasive Mucinous Adenocarcinoma of the Lung. J Thorac Oncol. 2016;11(4):496-503. Epub 20160130. doi: 10.1016/j.jtho.2016.01.010. PubMed PMID: 26829311.
46. Wang Z, Yuan X, Jiang G, Li Y, Yang F, Wang J, et al. Towards the molecular era of discriminating multiple lung cancers. EBioMedicine. 2023;90:104508. Epub 20230321. doi: 10.1016/j.ebiom.2023.104508. PubMed PMID: 36958271; PubMed Central PMCID: PMC10040518.
47. Liao CT, Wang HM, Chang JT, Ng SH, Hsueh C, Lee LY, et al. Analysis of risk factors for distant metastases in squamous cell carcinoma of the oral cavity. Cancer. 2007;110(7):1501-8. doi: 10.1002/cncr.22959. PubMed PMID: 17868119.
48. Pagedar NA, Jayawardena A, Charlton ME, Hoffman HT. Second Primary Lung Cancer After Head and Neck Cancer: Implications for Screening Computed Tomography. Ann Otol Rhinol Laryngol. 2015;124(10):765-9. Epub 20150416. doi: 10.1177/0003489415582259. PubMed PMID: 25881583; PubMed Central PMCID: PMC5296183.
49. Geurts TW, Nederlof PM, van den Brekel MW, van't Veer LJ, de Jong D, Hart AA, et al. Pulmonary squamous cell carcinoma following head and neck squamous cell carcinoma: metastasis or second primary? Clin Cancer Res. 2005;11(18):6608-14. doi: 10.1158/1078-0432.Ccr-05-0257. PubMed PMID: 16166439.
50. Bohnenberger H, Kaderali L, Ströbel P, Yepes D, Plessmann U, Dharia NV, et al. Comparative proteomics reveals a diagnostic signature for pulmonary head-and-neck cancer metastasis. EMBO Mol Med. 2018;10(9). doi: 10.15252/emmm.201708428. PubMed PMID: 30097507; PubMed Central PMCID: PMC6127892.

APPENDICES

Supplementary Table 1. The criteria of confidence level

Confidence level	Criteria
Likely	Max(P(IPM), P(MPLC)) < 0.8
Probable	0.8 ≤ Max(P(IPM), P(MPLC)) ≤ 0.95
Confident	Max(P(IPM), P(MPLC)) > 0.95

IPM: intrapulmonary metastasis, MPLC: multiple primary lung cancer.

Supplementary Table 2. The previous genomic based methods.

Name of algorithm		Cochin (2019)	MSK (2019)	HAPLOX (2021)
Journal		Journal of Thoracic Oncology	Clinical Cancer Research	Frontiers in Oncology
Panel (number of genes)		Ion AmpliSeq Colon-Lung Cancer Research Panel v2 (22 genes)	MSK-impact (468 genes)	Haplo-X (605 genes)
Clinical		AIS, MIA - MPLC Time interval over 5 years – MPLC	Same morphology - IPM Different histology - MPLC AIS/MIA/Lepidic predominant – MPLC	Unutilized
Molecular[1]	Non shared	MPLC	MPLC	MPLC
	1 (Rare)[2]	IPM	IPM	
	1 (Frequent)[2]	MPLC (*EGFR* (del19, L858R), *KRAS* G12X)	MPLC	MPLC (TKI genes)
	2	IPM	IPM	MPLC
	3	IPM	IPM	IPM
	No mutation	Indeterminate	Indeterminate	Indeterminate

[1]Molecular classification is based on the number of shared mutation.
[2]The algorithms classified a shared mutation as either rare or frequent in non-small cell lung cancer when there was only one mutation involved.
AIS: adenocarcinoma in situ, MIA: minimally invasive adenocarcinoma, MPLC: multiple primary lung cancer, IPM: intrapulmonary metastasis, TKI: Tyrosine kinase inhibitor

Supplementary table 3. Driver gene status of tumors from the in-house patients (available tumor n=128)

Molecular status (%)	
EGFR	80 (62.5)
Exon 18	
G719X+L861Q	1 (0.8)
Exon 19	
Exon 19 deletion[1]	28 (21.9)
Exon 20	
E20 insertion	1 (0.8)
Exon 21	
L858R	48 (37.5)
L858R+T790M	1 (0.8)
L861Q	1 (0.8)
Wild	48 (37.5)
ALK	
Positive	1 (0.8)
Negative	127 (99.2)
ROS1	
Positive[1]	3 (2.4)
Negative	125 (97.6)

[1] 1 case is concomitant ROS1 rearrangement and EGFR Exon 19 deletion.

Supplementary table 4. Clinical and pathologic characteristics of the 12 in-house pateints and their tumor pairs

Patient	Tumor ID	Sex	Age (at first resection)	1st histology	2nd histology	Lepidic component	CHA[1]	MeTel[1]	Seq. panel[2]	Location	Time interval (m)	Size (cm)	N stage	Smoking status	P-Y	DFS (m)	Recur	OS	Death
1)		**Similar histology with a long time interval**																	
Pt 1	E1_12	F	56	Acinar	Amucinous	Y	IPM	IPM	WES	LUL	60.8	3	2	nonsmoker		0.6	Y	44.9	N
	E1_17			Acinar	Amucinous	N				RUL		1							
Pt 2	E3_09	F	71	Alepidic	A lepidic	Y	MPLC	IPM	WES	RML	75.8	1.7	0	nonsmoker		23.6	Y	66.8	N
	E3_16_1			Acinar	A papillary	Y				RLL		1.7							
	E3_16_2			Acinar	A lepidic	Y				RLL		1.2							
Pt 3	E4_13[3]	F	53	Alepidic	A acinar	Y	MPLC		WES	LUL	75.5	2.5	0	nonsmoker		24.1	N	24.1	N
	E4_19_1		59	Acinar	A papillary	Y	IPM	IPM	WES	LLL	0	0.9	0						
	E4_19_2			Acinar	A papillary	N				LLL		1.3							
Pt 4	E5_13[3]	M	61	Apapillary	A acinar	Y	MPLC		WES	RUL	67.7	2.7	0	current	25	31.4	N	31.4	N
	E5_19_1		66	Acinar	A papillary	Y	IPM	IPM	WES	LLL	0	2							
	E5_19_2			Acinar	A papillary	N				LLL		1.5							
Pt 5	E6_14	F	68	Apapillary	A acinar	N	IPM	IPM	WES	RUL	63.4	2.7	0	nonsmoker		21.4	N	21.4	N
	E6_19			Apapillary	A acinar	N				LUL		1.2							
Pt 7	P11_09	M	66	Alepidic	A acinar	Y	MPLC	IPM	TSO500	LUL	99.3	3	0	nonsmoker		-	-	34.1	NA
	P11_17			Alepidic	A acinar	Y				LLL		2	0						
Pt 8	P12_07	F	45	Amicropapillary	A papillary	N	IPM	IPM	TSO500	LUL	129.8	2	0	nonsmoker		32.9	N	32.9	N
	P12_18			Amicropapillary	A papillary	N				LLL		1	0						
2)		**Synchronous multiple squamous cell carcinoma**																	

Pt 6	E7_15_1	M	64	Squamous	Squamous	NA	IPM	MPLC	WES	LLL	0.1	2.5	0	exsmoker	30	71.5	N	71.5	N
	E7_15_2			Squamous	Squamous	NA				RUL		2.5							
3)	**Similar histology with discordant driver gene status[4]**																		
Pt 9	P13_17	M	63	Acinar	A papillary	N	IPM	IPM	TSO500	LUL	32.2	1.2	0	nonsmoker		32.1	N	32.1	N
	P13_18			Acinar	A papillary	N				LUL		0.7	0						
4)	**Ambiguous histology with identical driver gene status**																		
Pt 10	P14_14	F	55	A papillary	A acinar	N	IPM	IPM	TSO500	LUL	28.1	2.8	0	nonsmoker		66.2	N	66.2	N
	P14_16			A papillary	A acinar + micropapillary	N				LLL		1.3	0						
Pt 11	P15_16	F	74	Acinar	A solid	N	MPLC	IPM	TSO500	RUL	13.3	2.9	0	nonsmoker		43.0	N	43.0	N
	P15_17			A solid	A acinar	N				LLL		2.2	0						
Pt 12	P16_15_1	F	67	Acinar	A papillary	Y	MPLC	MPLC	TSO500	LUL	0.5	1.8	2	nonsmoker		9.9	Y	67.3	Y
	P16_15_2			A papillary	A micropapillary	Y				RML		1.9	2						

[1] classification result by suggested criteria.

[2] WES was done for patient 1-6 and TSO500 sequencing was done for patient 7-12.

[3] The initial tumors of patients 3 and 4 (E4_13 and E5_13) were excluded due to low sequencing quality, hence the classification comparison was conducted on two later tumors (E4_19_1 and E4_19_2) (E5_19_1 and E5_19_2) from each patient.

[4] Patient 9 was initially diagnosed with multiple tumors showing discordant driver gene statuses, specifically EGFR E19 deletion and wild type. However, TSO500 sequencing conducted in this study revealed that the previously identified wild type was actually a false-negative for the identical EGFR E19 deletion status.

Pt: patient, CHA: Comprehensive histologic assessment, Seq.: sequencing, P-Y: pack-year, DFS: disease-free survival, OS: overall survival, F: female, M: male, A: adenocarcinoma, MPLC: multiple primary lung cancer, IPM: intrapulmonary metastasis, WES: whole exome sequencing, LUL: left upper lobe, LLL: left lower lobe, RUL: right upper lobe, RML: right middle lobe, RLL: right lower lobe, NA: not available.

Supplementary Table 5. The performance of algorithms. (279 cases in 6 test dataset)

		Cochin		**MSK**		**HAPLOX**		**MeTel**	
		IPM	MPLC	IPM	MPLC	IPM	MPLC	IPM	MPLC
Final Classification	IPM	70	5	66	9	35	40	75	0
	MPLC	11	193	3	201	10	194	7	197
Error Rate (Training set excluded)	Kappa coefficient	5.73% (9.41)	0.86	4.30% (5.91)	0.89	17.92 % (21.93)	0.48	2.51%	0.94

IPM: intrapulmonary metastasis, MPLC: multiple primary lung cancer.

Supplementary Table 6. F-1 scores for IPM and MPLC of algorithms. (279 cases in 6 test dataset)

	Cochin		**MSK**		**HAPLOX**		**MeTel**	
	IPM	MPLC	IPM	MPLC	IPM	MPLC	IPM	MPLC
Precision	0.86	0.97	0.96	0.96	0.78	0.83	0.91	1.00
Recall	0.93	0.95	0.88	0.99	0.47	0.95	1.00	0.97
F1	0.90	0.96	0.92	0.97	0.58	0.89	0.96	0.98

IPM: intrapulmonary metastasis, MPLC: multiple primary lung cancer.

Supplemantary table 7. Clinical and pathologic characteristics of discordant cases between histologic prediction and MeTel analysis

Patient	Tumor ID	Sex	Age (at first resection)	Acinar	Papillary	Micropapillary	Solid	Complex glandular	Lepidic	Squamous	CHA	MeTel	Seq. panel	Location	TI (m)	Size (cm)	Nstage	LVI	STAS	Smoking status	P-Y	DFS (m)	Recur	OS	Death
Pt2	E3_09	F	71	60	-	-	-	-	40	-	MPLC	IPM	WES	RML	75.8	1.7	0	N	N	nonsmoker	-	23.6	Y	66.8	N
	E3_16_1			60	15	15	-	-	10	-				RLL		1.7	0	N	Y						
	E3_16_2			50	15	5	-	-	30	-				RLL		1.2	0	N	Y						
Pt6	E7_15_1	M	64	-	-	-	-	-	-	MD	IPM	MPLC	WES	LLL	0.1	2.5	-	N	Y	exsmoker	30	71.5	N	71.5	N
	E7_15_2			-	-	-	-	-	-	MD				RUL		2.5	0	N	N						
Pt7	P11_09	M	66	30	-	-	-	-	70	-	MPLC	IPM	TSO500	LUL	99.3	3	0	N	N	nonsmoker	-	-	-	34.1	NA
	P11_17			15	-	5	-	-	80	-				LLL		2	0	N	N						
Pt11	P15_16	F	74	60	-	10	30	-	-	-	MPLC	IPM	TSO500	RUL	13.3	2.9	0	N	Y	nonsmoker	-	43.0	N	43.0	N
	P15_17			-	-	-	80	20	-	-				LLL		2.2	0	N	N						

WES: whole-exome sequencing, TI: time interval, LVI: lymphovascular invasion, STAS: spread through air spaces, P-Y: pack-year, DFS: disease-free survival, OS: overall survival, Pt: patient, F: female, M: male, MPLC: multiple primary lung cancer, IPM: intrapulmonary metastasis, LUL: left upper lobe, LLL: left lower lobe, RUL: right upper lobe, RML: right middle lobe, RLL: right lower lobe, MD: moderate differntiated, NA: not available.

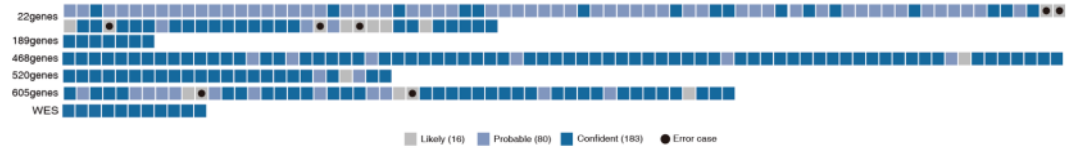
Supplementary Table 8. Mutation frequency showing the biggest difference by race.

	Mutation Frequency			
	Asian	Black	White	Total population
KRAS (p.G12C)	0.03	0.12	0.13	0.12
EGFR (p.L858R)	0.21	0.04	0.05	0.07

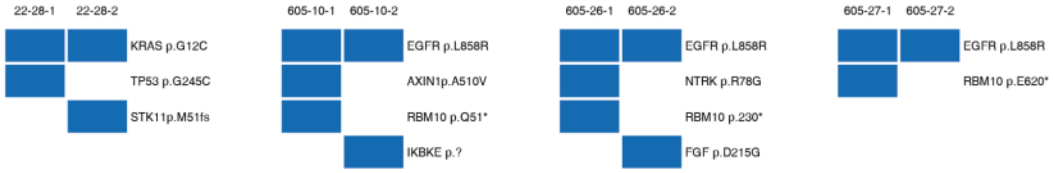
Supplementary table 9. Distribution of time interval between multiple tumors of 433 non-small cell lung cancer patients with multiple primary lung cancer and intrapulmonary metastasis at Yonsei University Severance Hospital. (2006 to 2020)

	IPM (N=156)			MPLC (N=277)		
Time interval	Frequency (N)	%	Cumultive %	Frequency (N)	%	Cumultive %
0-6M (synchronous)	106	67.9	67.9	204	73.6	73.6
6M-1Y	9	5.8	73.7	9	3.2	76.9
1Y-1Y6M	10	6.4	80.1	4	1.4	78.3
1Y6M-2Y	6	3.8	84.0	7	2.5	80.9
2Y-2Y6M	8	5.1	89.1	4	1.4	82.3
2Y6M-3Y	2	1.3	90.4	2	0.7	83.0
3Y-3Y6M	2	1.3	91.7	12	4.3	87.4
3Y6M-4Y	2	1.3	92.9	4	1.4	88.8
4Y-4Y6M	1	0.6	93.6	5	1.8	90.6
4Y6M-5Y	1	0.6	94.2	4	1.4	92.1
5Y-5Y6M	5	3.2	97.4	2	0.7	92.8
5Y6M-6Y				4	1.4	94.2
6Y-6Y6M	1	0.6	98.1	3	1.1	95.3
6Y6M-7Y				1	0.4	95.7
7Y-7Y6M				5	1.8	97.5
7Y6M-8Y						
8Y-8Y6M				2	0.7	98.2
8Y6M-9Y	1	0.6	98.7	1	0.4	98.6
9Y-9Y6M						
9Y6M-10Y	1	0.6	99.4	1	0.4	98.9
10Y-10Y6M	1	0.6	100.0	1	0.4	99.3
10Y6M-11Y				1	0.4	99.6
11Y-11Y6M						
11Y6M-12Y				1	0.4	100.0
Median TI (for cases TI>6M)	2Y-2Y6M			1Y-2Y		

MPLC: multiple primary lung cancer, IPM: intrapulmonary metastasis, M: month, Y: year, TI: time interval.



Supplementary figure 1. Confidence level and error cases of MeTel by sample. ‘Probable’ and ‘Confident’ showed 100% accuracy. All seven errors occurred only in ‘likely’ cases.



Supplementary figure 2. Changed cases when applying the ethnic specific frequency. One case (22-panel dataset from France) shared *KRAS*(p.G12C) and three cases (605-panel dataset from China) shared *EGFR*(p.L858R).

ABSTRACT(IN KOREAN)

독립된 다발성 폐암과 폐내 전이암의 감별에 대한 분자생물학적인 접근법

<지도교수 심 효 섭>

연세대학교 대학원 의학과

정 연 승

배경

폐암은 종종 다발성으로 발견되며 이는 독립된 다발성 폐암(Multiple primary lung cancer, MPLC) 또는 폐내 전이암(intrapulmonary metastasis, IPM)일 수 있다. 두 유형을 정확히 구분하는 것은 임상적으로 중요한 의미를 지니나 아직까지 그 진단 방법은 확립되지 않았다.

방법

본저에서는 유전체 정보를 기반으로 IPM과 MPLC를 구분하는 베이지안 확률 모델 (MeTel)을 개발하였다. 여섯 개의 코호트에서 다양한 크기의 패널(22개 유전자에서 전체 exome까지)로 시퀀싱된 279개의 샘플(75개의 IPM와 204개의 MPLC)을 대상으로 이전에 발표된 감별법들과 비교하였으며, 본원에서 감별이 모호했던 폐암 사례들을 모아 MeTel을 시행하고 기존 결과와 비교하였다.

결과

MeTel은 여섯 개 코호트를 대상으로 97.5%의 정확도를 보여 기존에 제시된 방법들(82.08-95.70%)을 능가하였으며, 패널의 크기에 잘 영향받지 않고 일관된 성능을 보였다. 나아가 MeTel의 분류를 통해, 본원에서 모호한 감별 결과를 얻었던 다발성 폐암 진단을 일부 수정할 수 있었다.

의의

본저에서는 IPM과 MPLC을 분류하는 새로운 유전체 기반 분류 알고리즘을 제시하고 그 정확성을 검증하였으며, 실제 임상 사례에 대한 활용성을 보였다.

핵심되는 말 : 다발성 폐암, 독립된 다발성 폐암, 폐내 전이암, 비소세포 폐암, 베이지안 확률 모델