JAMA | Original Investigation

Association of Patient Characteristics and Tumor Genomics With Clinical Outcomes Among Patients With Non–Small Cell Lung Cancer Using a Clinicogenomic Database

Gaurav Singal, MD; Peter G. Miller, MD, PhD; Vineeta Agarwala, MD, PhD; Gerald Li, PhD; Gaurav Kaushik, PhD; Daniel Backenroth, PhD; Anala Gossai, PhD, MPH; Garrett M. Frampton, PhD; Aracelis Z. Torres, PhD, MPH; Erik M. Lehnert, PhD; David Bourque, BS; Claire O'Connell, BS; Bryan Bowser, BS; Thomas Caron, BS; Ezra Baydur, BS; Kathi Seidl-Rathkopf, PhD; Ivan Ivanov, MS; Garrett Alpha-Cobb, PhD; Ameet Guria, BS; Jie He, PhD; Shannon Frank, BS; Allen C. Nunnally, JD; Mark Bailey, MS; Ann Jaskiw, BS; Dana Feuchtbaum, BS, MBA; Nathan Nussbaum, MD; Amy P. Abernethy, MD, PhD; Vincent A. Miller, MD

IMPORTANCE Data sets linking comprehensive genomic profiling (CGP) to clinical outcomes may accelerate precision medicine.

OBJECTIVE To assess whether a database that combines EHR-derived clinical data with CGP can identify and extend associations in non–small cell lung cancer (NSCLC).

DESIGN, SETTING, AND PARTICIPANTS Clinical data from EHRs were linked with CGP results for 28 998 patients from 275 US oncology practices. Among 4064 patients with NSCLC, exploratory associations between tumor genomics and patient characteristics with clinical outcomes were conducted, with data obtained between January 1, 2011, and January 1, 2018.

EXPOSURES Tumor CGP, including presence of a driver alteration (a pathogenic or likely pathogenic alteration in a gene shown to drive tumor growth); tumor mutation burden (TMB), defined as the number of mutations per megabase; and clinical characteristics gathered from EHRs.

MAIN OUTCOMES AND MEASURES Overall survival (OS), time receiving therapy, maximal therapy response (as documented by the treating physician in the EHR), and clinical benefit rate (fraction of patients with stable disease, partial response, or complete response) to therapy.

RESULTS Among 4064 patients with NSCLC (median age, 66.0 years; 51.9% female), 3183 (78.3%) had a history of smoking, 3153 (77.6%) had nonsquamous cancer, and 871 (21.4%) had an alteration in *EGFR*, *ALK*, or *ROS1* (701 [17.2%] with *EGFR*, 128 [3.1%] with *ALK*, and 42 [1.0%] with *ROS1* alterations). There were 1946 deaths in 7 years. For patients with a driver alteration, improved OS was observed among those treated with (n = 575) vs not treated with (n = 560) targeted therapies (median, 18.6 months [95% CI, 15.2-21.7] vs 11.4 months [95% CI, 9.7-12.5] from advanced diagnosis; *P* < .001). TMB (in mutations/Mb) was significantly higher among smokers vs nonsmokers (8.7 [IQR, 4.4-14.8] vs 2.6 [IQR, 1.7-5.2]; *P* < .001) and significantly lower among patients with vs without an alteration in *EGFR* (3.5 [IQR, 1.76-6.1] vs 7.8 [IQR, 3.5-13.9]; *P* < .001), *ALK* (2.1 [IQR, 0.9-4.0] vs 7.0 [IQR, 3.5-13.0]; *P* < .001), *RET* (4.6 [IQR, 1.7-8.7] vs 7.0 [IQR, 2.6-13.0]; *P* = .004), or *ROS1* (4.0 [IQR, 1.2-9.6] vs 7.0 [IQR, 2.6-13.0]; *P* = .03). In patients treated with anti–PD-1/PD-L1 therapies (n = 1290, 31.7%), TMB of 20 or more was significantly associated with improved OS from therapy initiation (16.8 months [95% CI, 11.6-24.9] vs 8.5 months [95% CI, 7.6-9.7]; *P* < .001), longer time receiving therapy (7.8 months [95% CI, 5.5-11.1] vs 3.3 months [95% CI, 2.8-3.7]; *P* < .001), and increased clinical benefit rate (80.7% vs 56.7%; *P* < .001) vs TMB less than 20.

CONCLUSIONS AND RELEVANCE Among patients with NSCLC included in a longitudinal database of clinical data linked to CGP results from routine care, exploratory analyses replicated previously described associations between clinical and genomic characteristics, between driver mutations and response to targeted therapy, and between TMB and response to immunotherapy. These findings demonstrate the feasibility of creating a clinicogenomic database derived from routine clinical experience and provide support for further research and discovery evaluating this approach in oncology.

JAMA. 2019;321(14):1391-1399. doi:10.1001/jama.2019.3241
Corrected on February 4, 2020.

Editorial page 1359

Supplemental content

Author Affiliations: Author affiliations are listed at the end of this article.

Corresponding Author: Gaurav Singal, MD, Foundation Medicine, Inc, 150 Second St, Cambridge, MA 02140 (gsingal@foundationmedicine.com).

The passage of the US Food and Drug Administration's Breakthrough Designation Act in 2012 and development of targeted therapeutics for patients with cancer have accelerated access to life-prolonging and, in some cases, curative medications, often based on substantial effect sizes seen in smaller, earlier-phase clinical trials.[1] In 2016, the 21st Century Cures Act recognized the need for "real-world evidence" to further establish the benefit of these agents, understand their optimal use, and inform future trial design.[2] Population-based, research-grade, linked clinicogenomic data sets generated from routine clinical care could significantly accelerate the advancement of clinical practice and the development of novel therapeutics.

Most efforts to identify clinicogenomic associations currently rely on clinical trials,[3] single-institution series,[4] or national registries.[5] Multi-institutional efforts, such as the American Association for Cancer Research's Genomics Evidence Neoplasia Information Exchange, have provided valuable insights but favor academic practice settings, whereas 85% of US patients receive oncologic care in the community.[6] Further, these observational data sets often are limited by heterogeneous or absent molecular testing, minimal clinical annotations of therapy and response, and limited or sometimes absent longitudinal updates.

The purpose of this study was to evaluate whether a newly developed, continuously updating clinicogenomic database—which included patient data collected routinely as part of health care delivery, with clinical data linked with comprehensive genomic profiling (CGP) results—was representative of known features of patients with non–small cell lung cancer (NSCLC), could replicate well-established genomic correlations with clinical outcomes, and would be useful for novel hypothesis generation, in particular related to immunotherapy.

Key Points

Question Can clinical and genomic data obtained in routine clinical care be linked in a Health Insurance Portability and Accountability Act–compliant manner to yield clinically relevant insights?

Findings A deidentified database of 28 998 patients with cancer, approximately 85% of whom were treated in a community setting, was generated by linking electronic health record–derived longitudinal clinical data with comprehensive tumor genomic profiling. Analysis of 4064 patients with non–small cell lung cancer revealed clinical, genomic, and therapeutic associations that were consistent with prior reports and extended previous observations on evolving community practice patterns.

Meaning Using data obtained from routine clinical care to generate a validated, multi-institution clinicogenomic database is feasible and can yield novel, clinically meaningful insights.

Methods

Regulatory Compliance and Cohort Generation

Approval for this study was obtained from the New England Institutional Review Board prior to study conduct, and the requirement for participant informed consent was waived. De-identification of patient data was performed according to Health Insurance Portability and Accountability Act guidelines and confirmed by an external statistical review process. The cohort was generated by linking the Flatiron Health electronic health records (EHRs) database to the Foundation Medicine database of tumor sequencing results (eFigure 1A in the Supplement).[7] Strictly deidentified "tokens" (hashed patient identifiers derived from, but not including, protected health information) were deterministically generated from each patient's demographic data and overlapped by a third party (Management Science Associates). Deidentified clinical and genomic data were submitted independently to Management Science Associates for linkage using the tokens without accessing personal health information. The linked, deidentified database was transferred for analysis with new deidentified tokens replacing the original tokens, preventing relinking to internal identified data sets (eFigure 1B in the Supplement). While the clinicogenomic database is refreshed every 3 to 6 months, the analyses reported here were based on the data as of January 1, 2018.

An NSCLC subcohort was defined by identifying patients who had an *International Classification of Diseases* (*ICD*) code for lung cancer within their EHR (*ICD-9*: 162.x; *ICD-10*: C34.x or C39.9), who were seen at an oncology practice in the clinical database at least once after January 2011, had medical record–confirmed diagnosis of NSCLC after manual review of the EHR, and had commercial CGP testing on a tumor specimen that a pathologist determined to be NSCLC or cancer of unknown primary. Patients were included if their CGP sample date was no earlier than 1 month prior to clinical diagnosis of NSCLC per medical record review.

Tumor Genotyping and PD-L1 Determination

Targeted DNA sequencing was performed on formalin-fixed, paraffin-embedded tissue using the FoundationOne platform, which includes full exonic coverage of 395 genes and intronic analysis for rearrangements at a depth of 500-1000x in its most current iteration (87.7% of all samples; eTable 1 in the Supplement).[8] Point mutations, amplifications, deletions, duplications, insertions, splice site variants, and rearrangements were categorized as either known pathogenic, likely pathogenic, variants of unknown significance, or germline single-nucleotide polymorphisms.[9] A National Comprehensive Cancer Network (NCCN)-listed driver mutation was defined as a pathogenic alteration in *EGFR*, *ALK*, *ROS1*, *MET*, *BRAF*, *RET*, or *ERBB2*. Tumor mutational burden (TMB), a measure of the number of somatic mutations identified per megabase of DNA sequenced, was calculated for every patient in the cohort.[10] Patients were stratified into TMB-high (TMB-H, ≥20 mutations/Mb) and TMB-low/intermediate (TMB-L/I, <20 mutations/Mb) based on prior efforts to define TMB cutoffs for response to immunotherapy.[11]

Programmed death-ligand 1 (PD-L1) levels were extracted from EHRs, as expression of this protein on tumor cells impairs immune-mediated tumor killing and may be associated with clinical response to immunotherapy agents targeting this pathway. A sample was classified as PD-L1 negative when reported as "negative" in the EHR or when testing by the central laboratory conducting next-generation sequencing revealed less than 50% staining using the 22C3 PD-L1 antibody.

A sample was classified as PD-L1 positive when reported as "positive" in the EHR or when testing by the central laboratory conducting next-generation sequencing revealed 50% or greater staining using either the 22C3 or SP142 PD-L1 antibodies. All other cases were deemed indeterminate or unknown. PD-L1 testing was included in survival and time receiving therapy analyses if reported before or within 30 days of the index start date. If multiple samples matched these criteria, the PD-L1 test temporally closest to the index date was selected.

Clinical Data Extraction

The clinical database contains longitudinal structured and unstructured EHR data from 275 cancer clinics representing more than 2 million actively treated US patients with cancer (approximately 85% from community practices). Structured data were harmonized and normalized to a standard ontology (aligning, where possible, to existing standards such as Logical Observation Identifiers Names and Codes for laboratory test results and North American Association of Central Cancer Registries standards for data elements commonly reported in tumor registries). Additional data points of interest (eg, smoking status, date of advanced disease diagnosis, biomarker status, and dates of disease progression) were extracted from unstructured EHR-derived digital documents via manual review by trained medical record abstractors (clinical oncology nurses and tumor registrars, with oncologist oversight) who followed prespecified, standardized policies and procedures. Race/ethnicity, as recorded in the EHR as part of clinical care, was included in this study given previous associations with outcomes in NSCLC.

End Points

The main end points were overall survival (time until death or loss to follow-up), time receiving therapy (duration of exposure to the therapy of interest within the earliest line containing said treatment; for instance, if a patient received erlotinib as part of both the second and fourth lines of therapy, the duration of the second line of therapy would be considered for this analysis), maximal therapeutic response (MTR; maximum response to a given therapy recorded in the EHR), and clinical benefit rate (CBR; the fraction of patients experiencing stable disease, partial response, or complete response). For most analyses, only a single end point was studied based on the end point best described previously. For some analyses with multiple previously described end points, each end point was studied and reported.

Overall survival was derived from a recent mortality data set generated by combining multiple data sources and benchmarked against the National Death Index.[12] Specifically, 4 sources of data were combined to develop a death data set of high sensitivity and specificity relative to standard National Death Index data: (1) EHR structured data contained within the clinical database; (2) abstracted data from EHR unstructured documents (eg, end-of-treatment notes or condolence letters); (3) commercial death data purchased from a vendor; and (4) publicly available US mortality data from the Social Security Death Index. Time receiving therapy was derived from structured medication orders and administrations recorded in the EHR, as well as manual medical record review of notes describing therapeutic regimens. The MTR variable was derived from retrospective review of longitudinal physician assessments of radiographic response to a particular therapy as documented in the EHR (using the NSCLC clinical tumor response variable, version 1.0); given that routine clinical practice does not include RECIST measurements, detailed policies and procedures for manual medical record review were defined to capture this variable reproducibly. MTR was classified as complete response, partial response, stable disease, or progressive disease. CBR was defined as the fraction of patients experiencing stable disease, partial response, or complete response.

Statistical Analysis

All patients in the cohort were included in the descriptive statistics, whereas all other analyses, including overall survival, time receiving therapy, and MTR, were restricted to the patients with advanced-stage disease unless otherwise noted. Descriptive statistics included median and range for continuous variables, and percentages and frequencies for categorical variables. 95% CIs around binomial proportions were calculated using either the Wilson score or Clopper-Pearson exact interval. Boxplots show the median and interquartile range (IQR) on the box, with whiskers extending to the furthest data point within 1.5 times the IQR. The log-rank test was used to compare groups. Confidence intervals on median survival time differences were calculated by bootstrapping 500 replicates. Patients were excluded from specific analyses when any relevant data (eg, PD-L1 test results) were missing, which varied based on data required for any given analysis. Missingness for each clinical variable is described in the **Table** and in general was limited across the variables.

In addition, 87.7% of patients were tested on the latest version of the CGP baitset (T7), each of whom had complete information on all 395 genes in the panel; the remaining 12.3% had sequencing performed on baitsets that had most of the same genes. TMB was available for every sequenced case. To account for left truncation, patients were treated as at risk of death only after the later of their sequencing report date and their first visit in the analytic database after January 1, 2011, as both are requirements for inclusion in the cohort. To account for potentially conflicting censor dates across the 2 data sets, records for whom the medical record terminated before their sequencing test was reported were removed from survival analyses. To account for potential immortal time bias, exposure to therapy was treated as a time-varying covariate in relevant analyses. Unless specified, only known or likely pathogenic alterations were considered in analyses.

In comparisons of gene frequencies between groups, the Benjamini-Hochberg method was applied and both uncorrected and corrected *P* values are reported to account for multiple hypothesis testing. The proportionality of hazards assumption was tested for all Cox model analyses by plotting the Schoenfeld residuals and was not violated unless specifically noted. Multivariable models incorporated age, sex, race/ethnicity, smoking status, pathology, line of first exposure to anti-programmed cell death protein 1 (PD-1)/PD-L1 therapy, *EGFR* mutations and *ALK* fusions, TMB, and PD-L1 status. Statistical tests were 2-sided. *P* values of less than .05 were considered statistically significant. As numerous associations were

Table. Cohort Sociodemographic and Tumor Characteristics

Characteristic	No. (%)
No. of patients	4064
Age at advanced diagnosis (n = 3522; completeness = 100%), y[a]	
Median (IQR)	66.0 (58.0-73.0)
Sex (completeness = 100%)	
Male	1955 (48.1)
Female	2109 (51.9)
Smoking status (n = 4011; completeness = 98.7%)	
History of smoking	3183 (79.3)
No history of smoking	828 (20.6)
Race/ethnicity (n = 3547; completeness = 87.3%)[b]	
White	2816 (79.4)
Other	372 (10.5)
Black or African American	227 (6.4)
Asian	126 (3.6)
Hispanic or Latino	6 (0.2)
Vital status: has documented date of death	1946 (47.9)
Stage of disease at initial diagnosis (n = 3848; completeness = 94.7%)[c]	
0	1 (<1)
I	388 (10.1)
II	309 (8.0)
III	846 (22.0)
IV	2304 (59.9)
Advanced disease status (completeness = 100%)[a]	3522 (86.7)
Histologic subtype (completeness = 100%)	
Nonsquamous cell carcinoma	3153 (77.6)
Squamous cell carcinoma	726 (17.9)
NSCLC histology not otherwise specified	185 (4.6)
No. of lines of therapy received[d]	
1	1183 (29.1)
2	811 (20.0)
≥3	755 (18.6)
No line of therapy captured in database	1315 (32.4)
Follow-up from initial diagnosis, median (IQR), d	34.0 (12.0-71.0)
Time between advanced diagnosis and Foundation Medicine test dates (n = 3522), mo	
Median (IQR)	2.0 (1.0-11.0)

Abbreviations: IQR, interquartile range; NSCLC, non–small cell lung cancer.

[a] Advanced diagnosis is defined as stage IIIb/IV disease at initial diagnosis or recurrent or metastatic disease at any stage.

[b] Race/ethnicity data were collected when recorded in the electronic health record as part of routine clinical care. "Other" included a myriad of terms describing race/ethnicity that were entered into the electronic health records; thus, it cannot be mapped to more specific categories.

[c] Stage of NSCLC is defined by the American Joint Committee on Cancer staging system and stage data were collected when recorded in the electronic health record as part of routine clinical care.

[d] Line of therapy is defined as a particular treatment regimen administered within a course of care. For example, a patient who has received 2 lines of therapy will have received 2 treatment regimens.

explored in these analyses, the conclusions should be considered exploratory and interpreted as demonstrations of feasibility. Analyses were performed on the R software version 3.5.1 (R Foundation for Statistical Computing) and Python software version 3.6.5 (Python Software Foundation).

Results

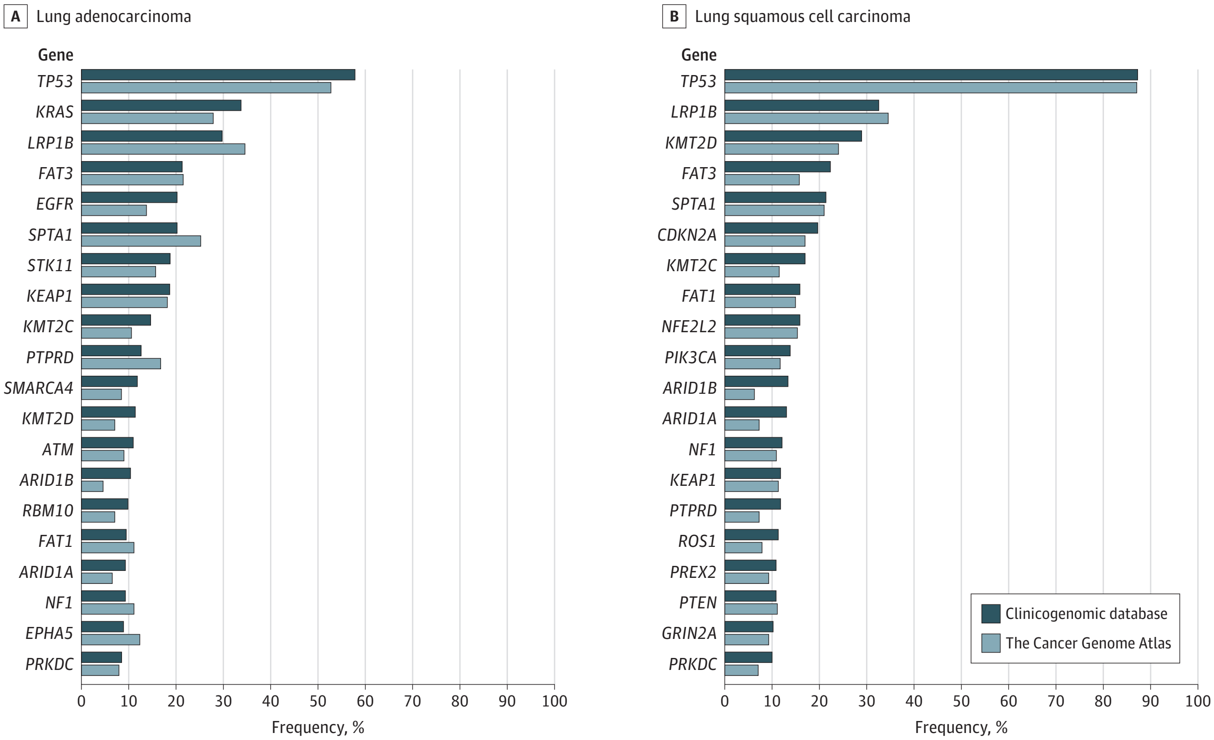
Clinicogenomic Features and Associations

The entire database included 28 998 patients across 38 tumor types, of which 4064 carried a diagnosis of NSCLC and were included in the cohort (eFigure 1C-D in the Supplement). The Table summarizes the clinical characteristics of the patients included in this analysis. The median age at diagnosis was 66.0 years (IQR, 58.0-73.0). Slightly more than half of patients were female (51.9%, n = 2109), most were white (79.4%, n = 2816), and most had a history of smoking (79.3%, n = 3183). A total of 77.6% (n = 3153) of the tumors were nonsquamous cell carcinoma, 17.9% (n = 726) were squamous cell carcinoma, and 4.6% (n = 185) not otherwise specified. Most patients had stage III (22.0%, n = 846) or IV (59.9%, n = 2304) disease at the time of initial diagnosis. At the time of this analysis, 86.7% of the patients had advanced disease, defined as stage IIIB/IV disease at diagnosis or recurrent or metastatic disease at any stage. Frequencies of data missingness are also reported in the Table (eg, for 1.3% of patients, smoking history information was unavailable in the health record).

The distribution of mutated genes was similar to previous descriptions from The Cancer Genome Atlas (data freeze 10), with a few notable exceptions (eg, point mutations and insertion/deletions in *EGFR*) (**Figure 1** and **Figure 2**; eFigure 2A-B in the Supplement).[13] TMB was significantly higher among smokers vs nonsmokers (median TMB [mutations/Mb], 8.7 [IQR, 4.4-14.8], n = 3183 vs 2.6 [IQR, 1.7-5.2], n = 828; $P < .001$) (eFigure 3A in the Supplement). Alterations in *EGFR*, *ALK*, *ROS1*, and *RET* were associated with significantly lower TMB than wild-type (WT) cases (median TMB [mutations/Mb] for *EGFR* mutant: 3.5 [IQR, 1.7-6.1], n = 701 vs WT: 7.8 [IQR, 3.5-13.9], n = 3363, $P < .001$; *ALK* mutant: 2.1 [IQR, 0.9-4.0], n = 128 vs WT: 7.0 [IQR, 3.5-13.0], n = 3936, $P < .001$; *ROS1* mutant: 4.0 [IQR, 1.2-9.6], n = 42 vs WT: 7.0 [IQR, 2.6-13.0], n = 4022; $P = .03$), and *RET* mutant: 4.6 [IQR, 1.7-8.7], n = 70 vs WT: 7.0 [IQR, 2.6-13.0], n = 3994; $P = .004$). Alterations in *PIK3CA* and *KRAS* were associated with significantly higher TMB (median TMB [mutations/Mb]: *PIK3CA* mutant: 8.7 [IQR, 3.6-13.9], n = 365 vs WT: 7.0 [IQR, 2.6-13.0], n = 3699, $P = .002$; *KRAS* mutant: 8.4 [IQR, 3.6-13.2], n = 1205 vs WT: 6.1 [IQR, 2.6-12.2], n = 2859, $P < .001$) (eFigure 3B-C in the Supplement). A co-mutation plot enabled visualization of sex, smoking status, pathology, stage, mutated genes, and TMB for each patient in the NSCLC cohort (eFigure 4 in the Supplement).

Associations between driver alterations and clinical characteristics were consistent with prior reports (eTable 2 in the Supplement). For example, *EGFR*, *ALK*, and *MET* alterations were enriched in nonsquamous pathologies (nonsquamous vs squamous: 91.6% vs 5.9%, $P < .001$ for *EGFR* mutant; 93% vs 3.9%, $P < .001$ for *ALK* mutant; and 86.4% vs 10%, $P = .009$ for *MET* mutant). Patients with *EGFR* mutations were more often Asian (mutant vs WT: 8.4% vs 2.0%; $P < .001$), female (62.8% vs 49.6%; $P < .001$), and never smokers (46.9% vs 14.8%; $P < .001$). Those with *ALK* translocations were more often younger (mutant vs WT: 58.0 vs 67.0 years old; $P < .001$) and never smokers (57.0% vs 19.2%; $P < .001$). Patients with *MET* alterations were more often older (mutant vs WT: 71.0 vs 66.0 years old; $P < .001$).

Figure 1. Comparison of Frequency of Mutations in Critical Genes in Non–Small Cell Lung Cancer Between the Clinicogenomic Database and The Cancer Genome Atlas



The frequency of short variant mutations (alterations that do not include translocations, large deletions, or copy number changes) identified in clinicogenomic database (CGDB) tumors analyzed on the most updated FoundationOne platform (T7 baitset; n = 2774 adenocarcinoma, n = 636 squamous cell carcinoma; eTable 1 in the Supplement) are shown in dark blue and among those in the The Cancer Genome Atlas (TCGA) (n = 567 adenocarcinoma, n = 495 squamous cell carcinoma) are shown in light blue for the 20 most commonly mutated genes. Adenocarcinomas and squamous cell cancers were analyzed separately given the well-established differences in mutational landscape between these tumor types. The genomic distribution and frequency of mutations were similar between the CGDB and TCGA in both adenocarcinoma and squamous cell lung cancers. For example, *TP53* was the most commonly mutated gene in both pathologies, *EGFR* mutations were more common in adenocarcinoma, and *CDKN2A* mutations were more common in squamous cell histology.

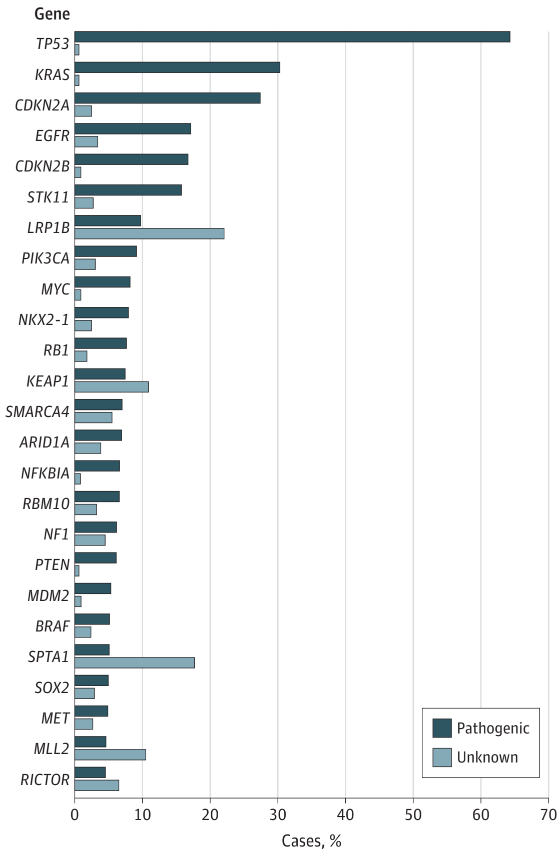
Prognostic Implications of Clinical and Genomic Features

All subsequent analyses were restricted to the patients in the cohort with advanced disease (n = 3522), unless otherwise noted. The median overall survival from advanced diagnosis across this cohort was 10.3 months (95% CI, 9.7-10.9) with a 5-year survival rate of 3.8% (eFigure 5A in the Supplement). Never smokers (n = 661) had longer overall survival from advanced diagnosis than those with a smoking history (n = 2564) (median overall survival, 15.2 months [95% CI, 12.9-18.4] vs 9.6 months [95% CI, 9.1-10.2]; difference, 5.5 [95% CI, 3.4-8.4]; *P* < .001), and patients with nonsquamous pathology (n = 2556) had longer overall survival than patients with squamous (n = 558) or not otherwise specified (n = 140) pathologies (median overall survival, 10.7 months [95% CI, 10.1-11.5] vs 9.6 months [95% CI, 8.7-11.3] vs 6.4 months [95% CI, 4.4-8.8], respectively; *P* < .001) (eFigure 5B-D in the Supplement). *TP53* (n = 2090 mutant, n = 1164 WT) and *RB1* (n = 241 mutant, n = 3013 WT) mutations were associated with shorter survival from advanced diagnosis (median overall survival, 9.6 months [95% CI, 9.1-10.4] vs 11.6 months [95% CI, 10.4-13.2]; difference, 1.9 [95% CI, 0.4-3.5], *P* < .001 and 8.6 months [95% CI, 7.3-11.2] vs 10.4 months [95% CI, 9.8-11.0]; difference, 1.7 [95% CI, −0.8 to 3.1], *P* = .01, respectively), whereas alterations in an NCCN-listed gene (n = 1134 mutant, n = 2120 WT) were associated with longer survival (median overall survival, 13.2 months [95% CI, 11.5-14.5] vs 9.4 months [95% CI, 8.8-10.0]; difference, 3.8 [95% CI, 2.1-5.1]; *P* < .001). The presence of a mutation in *KRAS* (n = 997 mutant, n = 2257 WT) was not associated with a difference in overall survival from advanced diagnosis (median overall survival, 9.6 months [95% CI, 8.7-10.7] vs 10.6 months [95% CI, 9.9-11.4]; difference, 1.0 [95% CI, −0.1 to 2.2]; *P* = .14) (**Figure 3**A; eFigure 6A in the Supplement).

Treatment Patterns and Genomic Variables Associated With Response to Targeted Therapy

Most patients with *EGFR*, *ALK*, and *ROS1* WT tumors were initially treated with a platinum-based chemotherapy regimen, whereas those with an *EGFR* or *ALK* alteration most commonly received a targeted anticancer agent (eFigure 7A in the Supplement). Among all patients with an advanced diagnosis who had an NCCN-driver alteration (n = 1260), 48.3% (n = 609) received NCCN-recommended therapy after the advanced diagnosis; 64.3% of patients with *EGFR* alterations (n = 405 of 630) and 70.1% of patients with *ALK* rearrangements (n = 75 of 107) received targeted therapies after advanced diagnosis.

Figure 2. Distribution of Mutations in Tumors for Patients With Non–Small Cell Lung Cancer (NSCLC) in the Cohort



To gain insight into the mutational landscape of the NSCLC clinicogenomic database cohort, alterations were identified for all patients tested on the most updated FoundationOne platform (n = 3564; eTable 1 in the Supplement). The alterations were then classified as likely impairing protein function and therefore pathogenic or of unknown significance using predefined algorithms (Methods section). The most commonly mutated gene across the cohort was *TP53*.

The most common second-line agent administered to those previously exposed to a platinum-based treatment was an anti-PD-1/PD-L1 agent (nivolumab, pembrolizumab, atezolizumab, durvalumab, or avelumab), a therapeutic class whose use increased over time during the observation period (eFigure 7B in the Supplement).

Among patients with a mutation in an NCCN-listed gene with evaluable overall survival, exposure to NCCN-directed targeted therapy treatment (n = 575 received, n = 560 did not receive) was associated with longer overall survival from advanced diagnosis (median overall survival, 18.6 months [95% CI, 15.2-21.7] vs 11.4 months [95% CI, 9.7-12.5]; difference, 7.1 [95% CI, 3.5-10.1]; *P* < .001) (Figure 3B). In particular, among patients with an *EGFR* alteration, treatment with an EGFR inhibitor (n = 380 received, n = 186 did not receive) was associated with longer overall survival from advanced diagnosis (median overall survival, 21.0 months [95% CI, 17.5-26.4] vs 13.3 months [95% CI, 10.8-14.6]; difference, 7.8 [95% CI, 4.7-13.9]; *P* < .001) (Figure 3C). Additionally, among all patients treated with an EGFR inhibitor, those with an *EGFR* alteration (n = 405) had a longer time receiving therapy (median time receiving therapy, 10.3 months [95% CI, 8.7-11.5] vs 2.8 months [95% CI, 2.4-3.7]; difference, 7.5 [95% CI, 5.6-8.7]; *P* < .001) and significantly higher MTR (CBR, 86.1% vs 50.9%; *P* < .001) compared with patients without an *EGFR* alteration (n = 148) (eFigure 7C and eTable 3 in the Supplement).

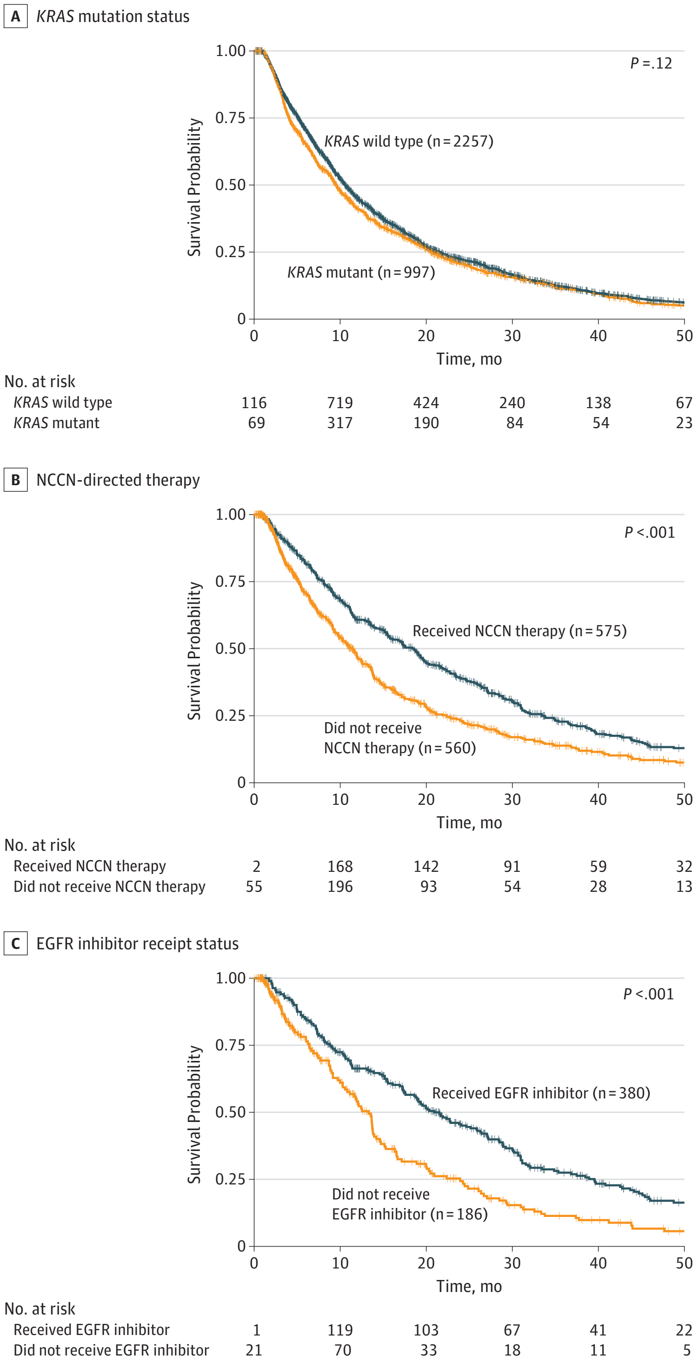
Immunotherapy, TMB, and PD-L1 Status

A total of 1290 patients received anti-PD-1/PD-L1 agents either as monotherapy or as part of a combination regimen. Compared with patients without anti-PD-1/PD-L1 therapy exposure (n = 2774), these patients had similar distributions of mutations and smoking history, with notable exception of significantly lower prevalence of alterations in *EGFR* (10.8% vs 20.4%; uncorrected and corrected *P* < .001) or *ALK* (1.3% vs 3.7%; uncorrected *P* < .001 and corrected *P* = .002) (eFigure 8A-C in the Supplement).

PD-L1 testing results were available for 1235 patients (11 commercial vendors, including Foundation Medicine). A total of 482 patients had positive or negative results and received anti-PD-1/PD-L1 therapy. Median TMB did not significantly differ in the PD-L1–positive (n = 289) vs -negative (n = 662) groups (median, 6.95 mutations/Mb [IQR, 3.5-14.0] in each group), consistent with prior reports (**Figure 4**A).[14] Bivariable analysis showed no association between PD-L1 status and time receiving therapy or overall survival among patients who received anti-PD-1/PD-L1 therapy (PD-L1–negative vs -positive median time receiving therapy, 3.7 months [95% CI, 2.8-5.2], n = 252 vs 5.5 months [95% CI, 4.3-6.5], n = 139; difference, 1.8 [95% CI, 0.1-3.2], *P* = .07; median overall survival, 10.1 [95% CI, 8.8-14], n = 248 vs 11.3 months [10.2-not reached], n = 137; difference, 1.2 [95% CI, −2.8 to 13.5], *P* = .31) (eFigure 9A-B in the Supplement).

Potential prognostic significance of TMB was assessed by analyzing patients with advanced disease who did not receive anti-PD-1/PD-L1 therapy. TMB-H was not associated with overall survival from advanced diagnosis in this population (TMB-H vs TMB-L/I median overall survival: 9.0 months [95% CI, 7.3-11.1], n = 226 vs 7.9 months [95% CI, 7.2-8.7], n = 1751; difference, 1.1 [95% CI, −0.89 to 2.8]; *P* = .11) (eFigure 9C in the Supplement). In the anti-PD-1/PD-L1–treated group, TMB-H was associated with significantly longer overall survival from start of treatment (median overall survival, 16.8 months [95% CI, 11.6-24.9], n = 161 vs 8.5 months [95% CI, 7.6-9.7], n = 1116; difference, 8.3 [95% CI, 2.1-15.6]; *P* < .001) and duration on anti-PD-1/PD-L1 therapy (median time receiving therapy, 7.8 months [95% CI, 5.5-11.1], n = 162 vs 3.3 months [95% CI, 2.8-3.7], n = 1127; difference, 4.5 [95% CI, 2.1-7.9]; *P* < .001) than TMB-L/I (Figure 4B; eFigure 9D in the Supplement). Similarly, the MTR to anti-PD-1/PD-L1 therapy was greater in the TMB-H group (n = 119) than in the TMB-L/I group (n = 789) (CBR, 80.7% vs 56.7%; *P* < .001) (eFigure 9E in the Supplement). In PD-L1–negative patients, TMB-H was not significantly associated with increased time receiving therapy or overall survival from start of therapy compared with TMB-L/I (median time receiving therapy, 11.3 months [95% CI, 2.8-not reached], n = 30 vs 3.5 months [95% CI, 2.8-4.8], n = 222; difference, 7.3 [95% CI, −1.0 to + ∞], *P* = .06 and median overall survival, 14.0 months [95% CI, 10.1-not reached], n = 29 vs 9.8 months [95% CI, 8.7-14],

Figure 3. Genomic Variables Associated With Survival and Therapy Response



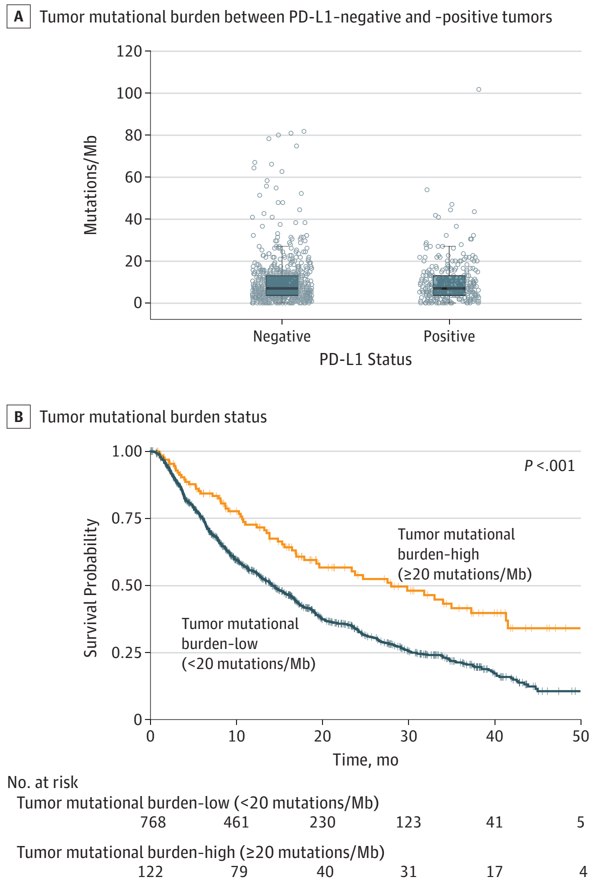
Overall survival from advanced diagnosis was determined and depicted. The 2 curves in each panel do not represent randomization but rather stratification based on *KRAS* mutation status for patients in the cohort (panel A, n = 3254; median observation time with the *KRAS* mutant: 15.8 months [interquartile range {IQR}, 7.7-27.7] and *KRAS* wild-type: 17.6 months [IQR, 7.6-30.6]), receipt of National Comprehensive Cancer Network (NCCN)–directed therapy among patients with a mutation outlined in the NCCN guidelines (panel B, n = 1135; median observation time for receipt: 16.9 months [IQR, 8.2-29.3] and no receipt: 17.6 months [IQR, 6.6-29.9]), or receipt of an epidermal growth factor receptor (EGFR) inhibitor among patients with an *EGFR* mutation (panel C, n = 566; median observation time for receipt: 16.6 months [IQR, 8.2-29.1] and no receipt: 21.8 months [IQR, 6.2-30.7]). Because patients are not intentionally randomized between groups, additional variables, such as physician practice patterns, may influence the between-group differences. The number of patients at risk initially increases because a subset of patients underwent comprehensive genomic profiling after their date of advanced diagnosis and therefore entered the risk pool at later times. The dynamic entry into the analytic cohort over the observation period also accounts for the difference between total numbers of patients in each cohort and the number at risk at any given time.

n = 219; difference, 4.2 [95% CI, −1.2 to + ∞], *P* = .10) (eFigure 9F-G in the Supplement). In a multivariable Cox proportional hazard analysis, higher TMB was associated with a significantly lower risk of discontinuing therapy (hazard ratio [HR], 0.78 for every additional 10 mutations/Mb [95% CI, 0.65-0.95]; *P* = .01), while receipt of anti-PD-1/PD-L1 therapy in second line or later (HR, 1.19 for every additional line after first [95% CI, 1.04-1.36]; *P* = .01) was associated with increased risk of discontinuing therapy (eTable 4 in the Supplement). PD-L1 status was not significantly associated with risk of discontinuing therapy (HR, 0.82 [95% CI, 0.62-1.12]; *P* = .21).

Discussion

Exploratory analyses of the NSCLC cohort of a clinicogenomic database generated by linking EHR-derived clinical data with CGP results obtained during routine clinical care from nearly

Figure 4. Immunotherapy and Tumor Mutational Burden



No. at risk

	0	10	20	30	40	50
Tumor mutational burden-low (<20 mutations/Mb)	768	461	230	123	41	5
Tumor mutational burden-high (≥20 mutations/Mb)	122	79	40	31	17	4

Because both programmed death-ligand 1 (PD-L1) expression level and tumor mutational burden (TMB) are potentially predictive biomarkers of immunotherapy response, the relationship between the 2 was investigated. A, There was no difference in the median TMB between the PD-L1–negative and -positive tumors. The heavy horizontal line is the median, the extremes of the box correspond to the 25th and 75th percentiles of the data, and the error bar extends to 1.5 times the interquartile range (IQR) from the edge of the box. B, Among patients who received PD-1/PD-L1–targeting therapy, those with a TMB (in mutations/Mb) greater than 20 (n = 161) had a longer overall survival from start of therapy than those with a TMB less than 20 (n = 1116) (median observation time for TMB ≥ 20: 14.3 months [IQR, 7.8-28.5] and for TMB < 20: 17.0 months [IQR, 7.9-30.0]).

29 000 patients representing more than 38 tumor types from more than 180 largely community-based oncology practices replicated previously described associations between clinical and genomic characteristics, driver mutations and response to targeted therapy, and TMB and response to immunotherapy.

The replication of clinicogenomic correlations previously identified through clinical trials, institutional studies, and academic sequencing efforts help establish the validity and applicability of this method. These findings are consistent with the results from the French Cooperative Thoracic Intergroup study of molecular alterations in advanced NSCLC.[5] Data sets derived from routine clinical practice may provide certain new or validated insights, such as the observation that treatment with an EGFR inhibitor in *EGFR* mutant disease was associated with prolonged survival, a finding that has been difficult to confirm in clinical trials due to crossover.

This data set further enabled the exploratory study of biomarkers associated with response to immunotherapy. Consistent with recent clinical trial data, TMB was found to be associated with both duration on anti-PD-1/PD-L1 therapy and survival from therapy start in bivariable and multivariable analyses, controlling for clinical and genomic confounders. Extending these findings, including the addition of multifactorial analyses with immune-associated genomic biomarkers, may help define additional and more robust variables associated with response to immunotherapies.

In addition, the use of large-scale clinicogenomic data sets can augment various stages of drug development. For example, a better understanding of the relationship between genomic drivers, patient information, and therapeutic response can inform target identification and focus discovery efforts on specific genomic subtypes, patients, or treatment modalities. Similarly, a deeper understanding of the genomic landscape across patient populations and how these patients respond to standard of care treatments may improve clinical trial design, including, for example, modeling expected performance of unexposed control populations for a novel therapy.

Limitations

This study has several limitations. First, a fundamental limitation of all overall survival analyses conducted using data collected from routine clinical practice is the quality and completeness of the underlying mortality data. While not every deceased patient's date of death was captured, and all biases in the distribution of this missingness were not known, the mortality data included in this clinicogenomic database had been previously validated as having high completeness and accuracy relative to gold-standard data sets.[12]

Second, the inclusion requirement of CGP testing results could create a survival bias; to mitigate this bias, left truncation was accounted for in the survival analyses. Third, bias could be introduced in the analysis of therapeutic exposures without randomization. To limit any potential immortal time bias, survival analyses comparing cohorts with differential treatment exposure incorporated therapy exposure as a time-varying covariate. Fourth, the requirement of CGP testing may introduce a selection bias for those with access to, and under the care of, physicians with distinct practice patterns. However, the data set included a broad set of patients including diversity in age, sex, and stage of disease.

Fifth, the time receiving therapy end point may not account for non-progression-related reasons for discontinuing therapy; efforts to integrate additional information (eg, radiology and pathology) may further refine this metric. Sixth, the MTR end point was limited by the physician's interpretation and documentation of the response; however, the ability to ascertain the interpretation of the treating clinician may reflect a clinically meaningful response not captured using other metrics. Seventh, the cohort was largely composed of patients with advanced-stage disease, a population in which CGP was more typically ordered. This needs to be considered when evaluating comparative results with other data sets, as the clinical and genomic landscape may change as the disease progresses.

Conclusions

Among patients with NSCLC included in a longitudinal database of clinical data linked to CGP results from routine care, exploratory analyses replicated previously described associations between clinical and genomic characteristics, between driver mutations and response to targeted therapy, and between TMB and response to immunotherapy. These findings demonstrate the feasibility of creating a clinicogenomic database derived from routine clinical experience and provide support for further research and discovery evaluating this approach in oncology.

ARTICLE INFORMATION

Accepted for Publication: March 6, 2019.

Correction: This article was corrected online February 4, 2020, to fix the height of the bars representing the percentage of cases with gene alterations of unknown significance in Figure 2.

Author Affiliations: Foundation Medicine Inc, Cambridge, Massachusetts (Singal, Li, Kaushik, Frampton, Lehnert, Bourque, Alpha-Cobb, Guria, He, Bailey, V. A. Miller); Brigham and Women's Hospital, Boston, Massachusetts (Singal); Department of Medical Oncology, Dana Farber Cancer Institute, Boston, Massachusetts (P. G. Miller); Flatiron Health Inc, New York, New York (Agarwala, Backenroth, Gossai, Torres, O'Connell, Bowser, Caron, Baydur, Seidl-Rathkopf, Ivanov, Frank, Jaskiw, Feuchtbaum, Nussbaum, Abernethy); Stanford University School of Medicine, Stanford, California (Agarwala); Voyager Therapeutics, Cambridge, Massachusetts (Nunnally); New York University School of Medicine, New York (Nussbaum); Now with the Food and Drug Administration, Silver Spring, Maryland (Abernethy).

Author Contributions: Dr P. G. Miller had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. Drs Singal, P. G. Miller, and Agarwala are co–first authors and contributed equally to this work.
Concept and design: Singal, P. G. Miller, Agarwala, Li, Kaushik, Frampton, He, Nunnally, Feuchtbaum, Nussbaum, Abernethy.
Acquisition, analysis, or interpretation of data: Singal, P. G. Miller, Agarwala, Li, Kaushik, Gossai, Backenroth, Frampton, Torres, Lehnert, Bourque, O'Connell, Bowser, Caron, Baydur, Seidl-Rathkopf, Ivanov, Alpha-Cobb, Guria, He, Frank, Bailey, Jaskiw, Abernethy, V. A. Miller.
Drafting of the manuscript: Singal, P. G. Miller, Agarwala, Kaushik, Gossai, Frampton, Frank, Bailey.
Critical revision of the manuscript for important intellectual content: Singal, P. G. Miller, Agarwala, Li, Kaushik, Backenroth, Frampton, Torres, Lehnert, Bourque, O'Connell, Bowser, Caron, Baydur, Seidl-Rathkopf, Ivanov, Alpha-Cobb, Guria, He, Nunnally, Jaskiw, Feuchtbaum, Nussbaum, Abernethy, V. A. Miller.
Statistical analysis: Singal, P. G. Miller, Li, Gossai, Backenroth, Frampton, Lehnert, O'Connell, He.
Obtained funding: Agarwala.
Administrative, technical, or material support: Singal, P. G. Miller, Agarwala, Torres, Bourque, O'Connell, Bowser, Caron, Baydur, Seidl-Rathkopf, Ivanov, Guria, Nunnally, Jaskiw, Feuchtbaum, Abernethy, V. A. Miller.
Supervision: Agarwala, Kaushik, O'Connell, Nunnally, Abernethy.

Conflict of Interest Disclosures: Dr P. G. Miller reported receiving personal fees from Foundation Medicine Inc during the conduct of the study. Drs Singal, Li, Kaushik, Frampton, Lehnert, Alpha-Cobb, He, and V. A. Miller; Mr Bourque; Mr Guria; and Mr Bailey are employees at Foundation Medicine. Drs Agarwala, Backenroth, Gossai, Torres, O'Connell, Seidl-Rathkopf, and Nussbaum; Mr Bowser; Mr Caron; Mr Baydur; Mr Ivanov; Ms Frank; Ms Jaskiw; and Ms Feuchtbaum are employees of Flatiron Health. Both Flatiron Health and Foundation Medicine are owned by the Roche Group. Dr Agarwala is also an employee of Google Ventures. Dr Seidl-Rathkopf also reported being an inventor for 2 patents for technology to generate and monitor cohort models for Flatiron Health. At the time of this work, Dr Abernethy was chief medical officer, chief scientific officer, and senior vice president of oncology at Flatiron Health, a member of the Roche Group, and had stock ownership in Roche. At that time, Dr Abernethy also declared the following: serving on the board of directors and stock ownership of athenahealth and CareDx; owner of Orange Leaf Associates LLC; senior advisor of Highlander Partners; advisor of SignalPath Research, RobinCare, and KelaHealth Inc; special advisor of The One Health Company; receiving honoraria from Roche/Genentech (<USD$10 000/year); and having a patent pending for a technology that facilitates the extraction of unstructured information from medical records. With the exception of Orange Leaf Associates LLC, all of these relationships ended on or before January 2019, predating her joining the US Food and Drug Administration (FDA). Since joining the FDA, Dr Abernethy has complied with applicable ethics laws. Dr V. A. Miller is chief medical officer of Foundation Medicine, a board member of Revolution Medicines from which he receives an honorarium and equity, and a patent holder for T790M for which he receives royalties from Memorial Sloan Kettering Cancer Center. No other disclosures were reported.

Funding/Support: Flatiron Health and Foundation Medicine provided the funding for this work. No external source of funding was obtained.

Role of the Funder/Sponsor: Representatives of Flatiron Health and Foundation Medicine were involved with the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript. Some authors are employees of the sponsors. The decision to submit this manuscript was made by the authors and not by the sponsor. The sponsors permitted submission of the publication without review of the content. The sponsors did not have the right to review or alter the content of the manuscript or revisions or to prevent publication.

Disclaimer: Dr Abernethy participated in this work prior to joining the FDA. This work and related conclusions reflect the independent work of study authors and does not necessarily represent the views of the FDA or US government.

REFERENCES

1. Vadola LA, Pond MA, Winter-Vann A, Whitsell. Faster approvals? trends in the use of FDA's expedited approval programs for oncology medications [published online May 30, 2017]. *J Clin Oncol*. doi:10.1200/JCO.2017.35.15_suppl.e18270

2. *21st Century Cures Act*. Pub L No. 114-255, 130 Stat 1033.

3. Jänne PA, Smith I, McWalter G, et al. Impact of KRAS codon subtypes from a randomised phase II trial of selumetinib plus docetaxel in KRAS mutant advanced non-small-cell lung cancer. *Br J Cancer*. 2015;113(2):199-203.

4. Zehir A, Benayed R, Shah RH, et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nat Med*. 2017;23(6):703-713.

5. Barlesi F, Mazieres J, Merlio JP, et al; Biomarkers France contributors. Routine molecular profiling of patients with advanced non-small-cell lung cancer. *Lancet*. 2016;387(10026):1415-1426.

6. Consortium APG; AACR Project GENIE Consortium. AACR Project GENIE. *Cancer Discov*. 2017;7(8):818-831.

7. Agarwala V, Khozin S, Singal G, et al. Real-world evidence in support of precision medicine. *Health Aff (Millwood)*. 2018;37(5):765-772.

8. Frampton GM, Fichtenholtz A, Otto GA, et al. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nat Biotechnol*. 2013;31(11):1023-1031.

9. Vignot S, Frampton GM, Soria JC, 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. *J Clin Oncol*. 2013;31(17):2167-2172.

10. Chalmers ZR, Connelly CF, Fabrizio D, et al. Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Med*. 2017;9(1):34.

11. Goodman AM, Kato S, Bazhenova L, et al. Tumor mutational burden as an independent predictor of response to immunotherapy in diverse cancers. *Mol Cancer Ther*. 2017;16(11):2598-2608.

12. Curtis MD, Griffith SD, Tucker M, et al. Development and validation of a high-quality composite real-world mortality endpoint. *Health Serv Res*. 2018;53(6):4460-4476. doi:10.1111/1475-6773.12872

13. Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*. 2014;511(7511):543-550.

14. Carbone DP, Reck M, Paz-Ares L, et al; CheckMate 026 Investigators. First-line nivolumab in stage IV or recurrent non-small-cell lung cancer. *N Engl J Med*. 2017;376(25):2415-2426.