cancers

MDPI

Article

Replication of Real-World Evidence in Oncology Using Electronic Health Record Data Extracted by Machine Learning

Corey M. Benedum [1], Arjun Sondhi [1], Erin Fidyk [1], Aaron B. Cohen [1,2], Sheila Nemeth [1], Blythe Adamson [1,3], Melissa Estévez [1,*,†] and Selen Bozkurt [1,†]

[1] Flatiron Health, Inc., 233 Spring Street, New York, NY 10003, USA; corey.benedum@flatiron.com (C.M.B.); arjun.sondhi@flatiron.com (A.S.); erin.fidyk@flatiron.com (E.F.); aaron.cohen@flatiron.com (A.B.C.); sheila.nemeth@flatiron.com (S.N.); badamson@flatiron.com (B.A.); selen.bozkurt@flatiron.com (S.B.)
[2] Department of Medicine, NYU Grossman School of Medicine, New York, NY 10016, USA
[3] Comparative Health Outcomes, Policy and Economics (CHOICE) Institute, University of Washington, Seattle, WA 98195, USA
* Correspondence: mestevez@flatiron.com
† These authors contributed equally to this work and should be considered as co-senior authors.

Simple Summary: Obtaining and structuring information about the characteristics, treatments, and outcomes of people living with cancer for research purposes is difficult and resource-intensive. Oftentimes, this information can only be found in electronic health records (EHRs). In response, researchers use natural language processing with machine learning (ML extraction) techniques to extract information at scale. This study evaluated the quality and fitness-for-use of EHR-derived oncology data curated using ML extraction, relative to the standard approach, abstraction by trained experts. Using patients with lung cancer from a real-world database, we performed replication analyses demonstrating common analyses conducted in observational research. Eligible patients were selected into biomarker- and treatment-defined cohorts, first with expert-abstracted then with ML-extracted data. The study's results and conclusions were similar regardless of the data curation method used. These results demonstrate that high-performance ML-extracted variables trained on expert-abstracted data can achieve similar results as when using abstracted data, unlocking the ability to perform oncology research at scale.

Abstract: Meaningful real-world evidence (RWE) generation requires unstructured data found in electronic health records (EHRs) which are often missing from administrative claims; however, obtaining relevant data from unstructured EHR sources is resource-intensive. In response, researchers are using natural language processing (NLP) with machine learning (ML) techniques (i.e., *ML extraction*) to extract real-world data (RWD) at scale. This study assessed the quality and fitness-for-use of EHR-derived oncology data curated using NLP with ML as compared to the reference standard of expert abstraction. Using a sample of 186,313 patients with lung cancer from a nationwide EHR-derived de-identified database, we performed a series of replication analyses demonstrating some common analyses conducted in retrospective observational research with complex EHR-derived data to generate evidence. Eligible patients were selected into biomarker- and treatment-defined cohorts, first with expert-abstracted then with ML-extracted data. We utilized the biomarker- and treatment-defined cohorts to perform analyses related to biomarker-associated survival and treatment comparative effectiveness, respectively. Across all analyses, the results differed by less than 8% between the data curation methods, and similar conclusions were reached. These results highlight that high-performance ML-extracted variables trained on expert-abstracted data can achieve similar results as when using abstracted data, unlocking the ability to perform oncology research at scale.

Keywords: electronic health records; machine learning; natural language processing; cancer; real-world data; artificial intelligence; quality; oncology; real-world evidence

check for updates

Citation: Benedum, C.M.; Sondhi, A.; Fidyk, E.; Cohen, A.B.; Nemeth, S.; Adamson, B.; Estévez, M.; Bozkurt, S. Replication of Real-World Evidence in Oncology Using Electronic Health Record Data Extracted by Machine Learning. *Cancers* **2023**, *15*, 1853. https://doi.org/10.3390/cancers15061853

Academic Editors: Andreas Stadlbauer, Anke Meyer-Baese and Max Zimmermann

Received: 27 January 2023
Revised: 7 March 2023
Accepted: 14 March 2023
Published: 20 March 2023

1. Introduction

The digitization of healthcare, driven in part by the Health Information Technology for Economic and Clinical Health (HITECH) Act signed into US law in 2009, has increased the availability of real-world data (RWD). Likewise, the demand for real-world evidence (RWE) to support comparative effectiveness research and better understand patient populations and clinical outcomes has grown [1–4]. Despite this growth in available patient data, 80% of RWD is in unstructured free-text and requires manual curation and processing to be usable for analysis purposes [4,5]. Valuable information regarding the characteristics, treatments, and outcomes of people living with cancer is found in unstructured text documents stored within electronic health records (EHRs). How to access and analyze this information at scale for RWE generation is a massive challenge. The standard method of data curation through expert human abstraction is resource-intensive and time-consuming, limiting the number of patients available for research purposes [5–7]. In response, natural language processing (NLP) with machine learning (ML) techniques (i.e., *ML extraction*) is increasingly being applied to EHR data for more efficient and scalable generation of RWD (Box 1). ML extraction techniques can learn and recognize language patterns to provide automated solutions for extracting clinically relevant information, thereby enabling research and RWE generation at scale [8] (Figure 1). As researchers seek to understand smaller and more niche patient populations and stay on top of rapidly evolving standards of care, the need to generate RWD quickly and for more patients is becoming increasingly important. By automatically processing free text to extract clinical information, ML extraction can generate RWD at a speed and scale that far exceeds manual data curation and thereby meet the evolving needs of clinical and health outcomes research. For example, ML extraction can scan an enormous population, searching for rare patient characteristics buried in unstructured EHR data sources to select niche populations and unlock larger cohort sizes than would be feasible with expert abstraction.

Box 1. Defining key terms.

Natural language processing (NLP) is a tool used to enable computers to analyze, understand, derive meaning from, and make use of human language. Often, NLP is applied to identify and extract relevant information from unstructured data. The output of this document processing is a set of features which capture document structure, chronology, and key clinical terms or phrases. These features can then serve as the inputs for a machine learning model.
Machine learning (ML) can also be used to perform NLP to extract data from unstructured sources. ML models are designed to learn to perform tasks without being explicitly programmed to do so. For example, a ML model can be trained to learn what keywords or phrases found in a patient's clinical documents are associated with a variable of interest.

Most regulatory guidance related to ML has primarily focused on evaluating ML models and software as a medical device [9]. Until recently, there had been limited regulatory guidance regarding the best practices for evaluating ML-extracted RWD, aside from an overarching agreement on the need for transparent methods and processes. Both the UK's NICE RWE framework and FDA's RWE guidance ultimately aim to deliver on this by improving RWE quality through detailed guidance on what constitutes RWD, data curation, and analysis reporting standards, measuring quality and addressing limitations such as missing data or information bias [10–12]. While there is growing attention and guidance around RWD at large, there remains a gap regarding the evaluation of the quality and performance of ML-extracted RWD.

In response to this gap, we previously developed a research-centric evaluation framework to evaluate ML-extracted RWD and provide insights on model performance, strengths and limitations, and fitness-for-use [6]. This framework primarily focuses on evaluating a single ML-extracted variable, independent of the output of other ML extraction models. Univariable analyses include characterizing the model's overall performance and performance stratified by key patient characteristics, quantitative error analysis, where the characteristics of correctly and incorrectly extracted patients are compared, to understand the potential for systematic bias due to model errors, and finally a comparison of the outcomes between cohorts selected by the model as compared to expert abstraction. While understanding the quality of the extracted data for individual variables is important, univariable evaluations cannot describe how model errors may interact together and potentially introduce bias when multiple ML-extracted variables are used in combination for research purposes (e.g., selection bias due to poor model performance in select sub-groups or information bias, resulting in shifts in covariate distributions) [13,14]. As such, replication analyses leveraging datasets containing several ML-extracted variables are integral to understanding the reproducibility of analytic results and scientific conclusions when using data curated via ML extraction versus expert abstraction.

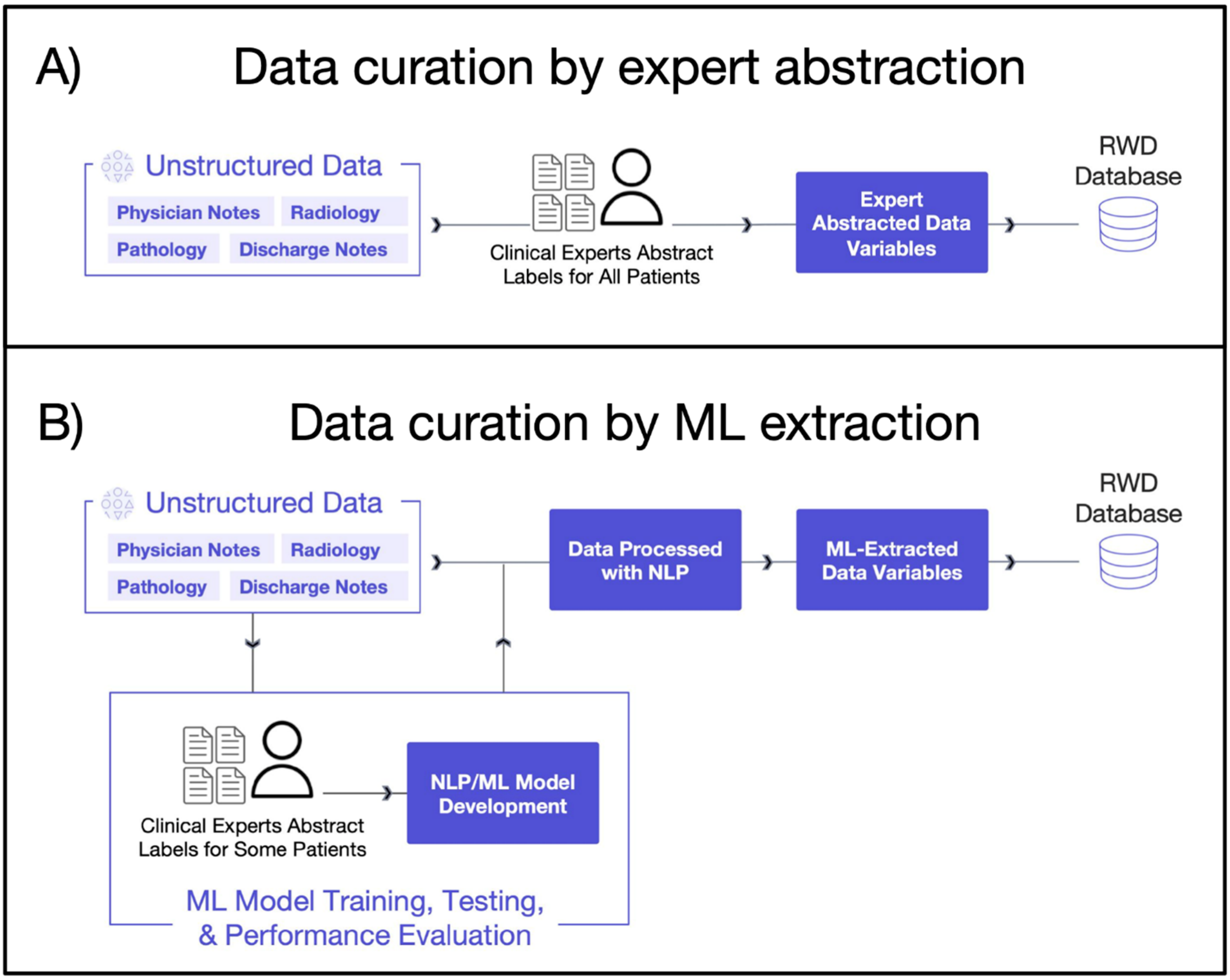
A)
Data curation by expert abstraction
Unstructured Data
Physician Notes
Radiology
Pathology
Discharge Notes
Clinical Experts Abstract Labels for All Patients
Expert Abstracted Data Variables
RWD Database
B)
Data curation by ML extraction
Unstructured Data
Physician Notes
Radiology
Pathology
Discharge Notes
Data Processed with NLP
ML-Extracted Data Variables
RWD Database
Clinical Experts Abstract Labels for Some Patients
NLP/ML Model Development
ML Model Training, Testing, & Performance Evaluation

Figure 1. Conceptual diagram of EHR data curation highlighting approaches to define variables. Abbreviations: ML: machine learning; NLP: natural language processing; RWD: real-world data. (Panel (**A**)): Unstructured data are reviewed by trained clinical abstractors to collect relevant data from patients' charts. (Panel (**B**)): Process for developing models and extracting information from unstructured data sources from the patient's chart.

We identified common archetypes describing how EHR-derived data are used in observational research. These archetypes include but are not limited to: (1) defining baseline characteristics, (2) describing the natural history of disease, (3) balancing populations, and (4) measuring treatment comparative effectiveness. For this study, we designed example oncology retrospective studies for each archetype that require information from both unstructured and structured complex EHR data, and assessed whether the use of ML-extracted data leads to the same analytic conclusions when used in place of expert-abstracted data for each. The retrospective studies we designed include two study populations: a biomarker-defined cohort and a treatment-defined cohort. These populations were selected because the ability of ML-extracted data to select these cohorts unlocks the ability to perform comparative effectiveness research and understand outcomes, including in rare populations that can benefit from targeted therapy. For the biomarker-defined cohort, we selected patients with a ROS1 rearrangement (in addition to ROS1-negative patients) to evaluate the ability of ML-extracted data to select a cohort with a low prevalence. For the treatment-defined cohort, we chose patients who received first line (1L) treatment with either bevacizumab–carboplatin–paclitaxel (BCP) or carboplatin–paclitaxel (CP). Since the goal of this study is to compare ML extraction with expert abstraction and not to contribute to the scientific understanding of cancer treatment, we intentionally selected populations that are well established in the literature.

2. Materials and Methods

We developed a series of retrospective study replications in non-small cell lung cancer (NSCLC) to compare conclusions based on ML-extracted data relative to expert-abstracted data. Two research questions were defined to illustrate the common archetypes for RWD use cases:

1. *What is the relationship between a rare cancer biomarker alteration and patient survival?*
2. *What is the comparative effectiveness of two cancer treatment regimens?*

For each research question, we defined an analytic cohort and selected patients who met the cohort eligibility criteria using variables defined with expert-abstracted and structured data (i.e., *abstracted cohort*) and subsequently those who met the cohort selection criteria using ML-extracted and structured data (i.e., *ML-extracted cohort*). We then performed analyses related to each archetype using the *abstracted cohort* and the *ML-extracted cohort*. Results and conclusions based upon these results were compared between data curation approaches.

2.1. Data Source

The data used to generate the results of this study were obtained from Flatiron Health's US-nationwide EHR-derived database, which includes longitudinal de-identified data from ~280 cancer practices (approximately 800 distinct sites of care) curated via technology-enabled abstraction [5,15]. The distribution of patients across community and academic practices largely reflects patterns of care in the US, where most patients are treated in community clinics, but this can vary for each disease. Mortality information is captured via a composite variable that uses multiple data sources (structured and unstructured EHR-derived content, commercial sources, Social Security Death Index) and is benchmarked against the National Death Index data as the gold standard [16]. We obtained the key analysis variables from both structured and unstructured (e.g., physician notes, pathology reports, discharge summaries) data sources in the patient's EHR (Table 1). A data cutoff date of 30 November 2022 was used, meaning that all information recorded into the EHR through 31 October 2022 would be included. Unstructured data were then curated by both expert clinical abstractors and ML models (Figure 1).

2.1.1. Expert Abstraction

All manual abstraction of unstructured information is carried out by trained abstractors (i.e., clinical oncology nurses or tumor registrars). Clinically relevant details are abstracted from relevant forms of clinical documentation available in the EHR, including clinic visit notes, radiology reports, pathology reports, etc. Abstractors are trained to identify and extract relevant information by following policies and procedures that are tested and optimized for reliability and reproducibility through iterative processes, and oversight is provided by oncology clinicians. The database undergoes continuous audit procedures to monitor abstractor performance while proprietary technology links each curated data variable to its source documentation within the EHR, enabling a subsequent review when necessary. Further, these data undergo quality assurance/quality control procedures to ensure data conformance, plausibility, and consistency. At the individual patient level, this approach provides a recent and robust longitudinal view into the clinical course, capturing new clinical information as it is documented within the EHR.

Table 1. Study variables and EHR data source.

EHR Source Information Type	Variables Needed for Analysis	Curation Approaches
Structured data (e.g., date of birth)	1. Diagnoses (i.e., ICD codes) 2. Gender 3. Birth year 4. Race 5. Ethnicity 6. Practice type 7. ECOG performance status 8. Medication order date 9. Medication administration date 10. Visit date 11. Mortality date [a]	Transformation, harmonization, and deduplication
Unstructured data (e.g., clinic notes, PDF lab reports, radiology images, etc.)	1. NSCLC diagnosis 2. NSCLC diagnosis date 3. Advanced NSCLC diagnosis 4. Advanced NSCLC diagnosis date 5. *ROS1* test result 6. *ROS1* test date 7. *ALK* test result 8. *BRAF* test result 9. *EGFR* test result 10. *ALK* test date 11. *BRAF* test date 12. *EGFR* test date 13. PD-L1 percent staining 14. PD-L1 test result date 15. Group stage 16. Histology 17. Line of therapy [b] 18. Line of therapy start date [b]	Expert abstraction OR ML-extraction

Abbreviations: *ALK*: anaplastic lymphoma kinase; *EGFR*: epidermal growth factor receptor; ECOG: Eastern Cooperative Oncology Group; ML: machine learning; NSCLC: non-small cell lung cancer; PD-L1: programmed death-ligand 1; [a] mortality date is a composite variable based on multiple data sources (structured and unstructured EHR data, commercial sources, and Social Security Death Index) [16]. *ML extraction was not used to define this variable.* [b] Line of therapy and line of therapy date are a derived variable based on both structured and unstructured data inputs.

2.1.2. Machine Learning Extraction

A multi-disciplinary ML team (including oncology clinicians, engineers, quantitative scientists, and other experts) developed a set of nine distinct models for key analysis variables (Table 1) that would not otherwise be available in structured EHR or claims data. Each of the 18 variables has been extracted through NLP of clinical notes, followed by an advanced ML or deep learning model, including LSTM and XGBoost, after undergoing a rigorous development, validation, and testing process that aligns with the data and the model's objectives. Model details, such as how they were developed, have been previously described [17]. Briefly, models are trained on the data labeled by expert abstraction to recognize, interpret, and curate free text into structured variable values in order to mimic the abstraction process. Models used between 35,710 and 211,581 expert-abstracted labels for training, validation, and testing, depending on the variable. The trained models then extracted relevant information using the same clinical documents available to the expert abstractors. In this context, NLP is used to identify sentences in relevant unstructured

EHR documents (e.g., oncology visit notes, lab reports, etc.) that contain a match to one of the clinical terms or phrases. These sentences are then transformed into a mathematical representation that the model can interpret. Individual models used in this study were evaluated with the research-centric evaluation framework developed by Estevez et al. [6]. Each model's performance was evaluated using a test set of over 3000 unique lung cancer patients.

2.2. Study Population

We selected a population of patients, sampled from the study database, with the following inclusion criteria: a lung cancer ICD code (ICD-9 162.x or ICD-10 C34x or C39.9) and at least two unique-date clinic encounters documented in the EHR in the study database (reflected by records of vital signs, treatment administration, and/or laboratory tests) on or after 1 January 2011. Among this population, we applied study eligibility criteria for each research question and selected two distinct cohorts for analysis. Some of the cohort selection criteria used variables that were defined by expert-abstracted and structured data and then replicated using ML-extracted and structured data (i.e., the *abstracted* and *ML-extracted cohort*, respectively). The selection of patients for the biomarker-defined population and treatment-defined population is described in Figure 2.

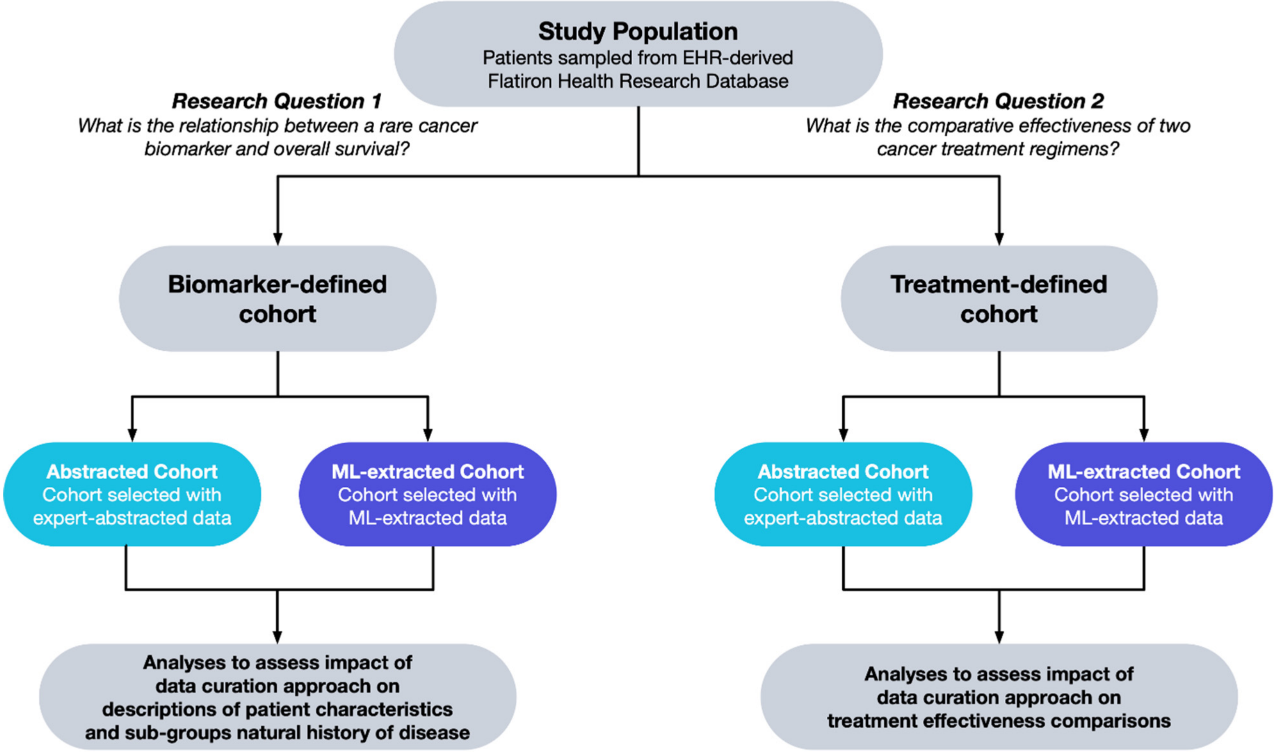
Study Population
Patients sampled from EHR-derived
Flatiron Health Research Database
Research Question 1
What is the relationship between a rare cancer biomarker and overall survival?
Research Question 2
What is the comparative effectiveness of two cancer treatment regimens?
Biomarker-defined cohort
Treatment-defined cohort
Abstracted Cohort
Cohort selected with expert-abstracted data
ML-extracted Cohort
Cohort selected with ML-extracted data
Abstracted Cohort
Cohort selected with expert-abstracted data
ML-extracted Cohort
Cohort selected with ML-extracted data
Analyses to assess impact of data curation approach on descriptions of patient characteristics and sub-groups natural history of disease
Analyses to assess impact of data curation approach on treatment effectiveness comparisons

Figure 2. Data curation approach for replication analyses. Abbreviations: ML: machine learning.

2.2.1. Biomarker-Defined Cohort

To answer the research question of survival by biomarker status, we selected patients diagnosed with NSCLC between 1 January 2011 and 31 October 2022 having advanced disease, defined as being either stage IIIB or higher upon diagnosis, or those who had earlier stage disease with subsequent development of recurrent or metastatic disease, and either (1) ever-positive status for a *ROS1* rearrangement after NSCLC diagnosis or (2) only negative status for a *ROS1* rearrangement in addition to a never-positive status for *ALK* (anaplastic lymphoma kinase) rearrangement, *BRAF* mutation, and *EGFR* (epidermal growth factor receptor) mutation, after NSCLC diagnosis. Patients were excluded in this cohort if they did not have a test result or only an unknown test result for the biomarker of interest, *ROS1*.

2.2.2. Treatment-Defined Cohort

To answer the research question comparing the effectiveness of cancer treatment regimens, we selected a cohort of patients diagnosed with de novo stage IV non-squamous NSCLC between 1 January 2011 and 31 October 2022 who received 1L treatment with either BCP or CP. Additional eligibility criteria were applied for adequate organ function as measured by lab test results and the ECOG performance status (eligibility criteria defined in the Supplementary Materials Table S3).

2.3. Statistical Analysis

Statistical analyses were designed to be illustrative in demonstrating the previously defined common research archetypes. These include:

1. Defining baseline characteristics;
2. Describing natural history of disease in biomarker sub-groups;
3. Balancing populations;
4. Measuring treatment comparative effectiveness.

All analyses were first performed using the abstracted cohort and data curated by expert abstractors. Using identical methods and code, we executed the same analyses using the ML-extracted cohort and data curated by ML models. The results were then compared between data curation approaches. We used the biomarker-defined cohort to evaluate the reproducibility of archetypes 1 and 2 and the treatment-defined cohort to evaluate the reproducibility of archetypes 3 and 4.

2.3.1. Defining Baseline Characteristics

We summarized select patient demographics and clinical characteristics, obtained from both structured and unstructured data sources, with descriptive statistics (i.e., median and IQR for continuous variables; n and percent for categorical variables), stratified by *ROS1* rearrangement status. Using the absolute standardized mean difference (aSMD), we compared the distribution of these characteristics within the *ROS1* rearrangement strata between the abstracted and ML-extracted cohorts. Comparisons where the aSMD was less than 0.1 were considered negligible [18]. Evaluated characteristics that were curated by ML models in the replication include: cancer histology, age at advanced diagnosis, advanced diagnosis year, group stage at NSCLC diagnosis, smoking status, treatment type received, and ever positive for: *ALK* rearrangement, *BRAF* mutation, *EGFR* mutation, or PD-L1 (programmed death-ligand 1) expression.

2.3.2. Natural History of Disease in Biomarker Sub-Groups

The real-world overall survival (rwOS) was calculated as the time from advanced diagnosis date to death, using a risk set-adjusted Kaplan–Meier estimator, so that patients are only counted at risk for death once the patient has met the cohort entry criteria [19,20]. The results are stratified by the *ROS1* result (*positive* or *negative*). We compared the rwOS of patients who were *ROS1*-positive versus -negative using univariate and matched and adjusted Cox proportional hazards models to estimate the hazard ratio (HR) and 95% confidence interval (CI). The Supplementary Materials describe further details on the univariate and matched models, such as modeling and matching procedures, covariates statistically controlled for, and a robustness check to evaluate an alternative covariate selection approach for the matched model.

2.3.3. Balancing Populations

To balance the baseline characteristics of patients who received different treatment regimens in the treatment-defined cohort, we fit a propensity score model [18] that included the treatment start year, age, sex, race/ethnicity, smoking status, and biomarker positivity status [21]. We assigned inverse probability weights (IPW) to weight each treatment arm.

2.3.4. Comparative Effectiveness Analysis

To estimate the average treatment effect (ATE) parameter, we used the IPW weighted population from the treatment-comparison cohort. We fit a Cox proportional hazards model with a treatment group indicator (*BCP*, *CP*). We summarized the comparison of rwOS between treatment groups using the estimated HR and 95% CI.

We performed all analyses using the R programming language version 4.1.3 [22]. Institutional Review Board approval of the study protocol was obtained prior to the study's conduct, and included a waiver for informed consent.

3. Results

3.1. Biomarker-Defined Cohort

The selection of the biomarker-defined cohort included 27,478 patients when using data curated by expert abstraction and 29,586 patients when using data curated by ML extraction. Patient attrition for this cohort when using expert-abstracted and ML-extracted data is described in Supplementary Materials Table S1.

3.1.1. Defining Baseline Characteristics

There were no clinically meaningful differences in the distribution of baseline characteristics for the patients selected using expert-abstracted compared to ML-extracted variables (Table 2). The prevalence of a positive biomarker test result for *ROS1* rearrangement was 1.27% (abstracted cohort) and 1.24% (ML-extracted cohort). Among biomarker-positive patients, there were small differences (aSMD < 0.2) between the abstracted and ML-extracted cohorts in the characteristics of diagnosis year, disease stage, ECOG performance status, and treatments. There were no differences among patients who were biomarker-negative.

Table 2. Baseline characteristics of patients, by biomarker result and data curation approach.

	ROS1-Positive			**ROS1-Negative**		
	Abstracted Cohort	**ML-Extracted Cohort**	**aSMD**	**Abstracted Cohort**	**ML-Extracted Cohort**	**aSMD**
N	349	367		27,478	29,219	
Practice Type, n (%)			0.02			0.01
Academic	94 (26.9%)	102 (27.8%)		3907 (14.2%)	4032 (13.8%)	
Community	255 (73.1%)	265 (72.2%)		23,571 (85.8%)	25,187 (86.2%)	
Gender, n (%)			0.06			0.00
Female	217 (62.2%)	218 (59.4%)		12,966 (47.2%)	13,790 (47.2%)	
Male	132 (37.8%)	149 (40.6%)		14,510 (52.8%)	15,427 (52.8%)	
Race/ethnicity, n (%)			0.03			0.01
Black or African American	38 (10.9%)	42 (11.4%)		2419 (8.8%)	2568 (8.8%)	
Other race [a]	64 (18.3%)	70 (19.1%)		3637 (13.2%)	3901 (13.4%)	
Unknown	32 (9.2%)	35 (9.5%)		2794 (10.2%)	3009 (10.3%)	
White	215 (61.6%)	220 (59.9%)		18,628 (67.8%)	19,741 (67.6%)	
Age at advanced diagnosis, median [IQR]	65 (55, 75)	65 (54, 74)	0.02	69 (62, 76)	69 (62, 76)	0.00
Advanced diagnosis year, n (%)			0.16			0.04
2011	3 (0.9%)	4 (1.1%)		95 (0.3%)	105 (0.4%)	
2012	10 (2.9%)	7 (1.9%)		253 (0.9%)	272 (0.9%)	
2013	12 (3.4%)	16 (4.4%)		638 (2.3%)	676 (2.3%)	
2014	18 (5.2%)	18 (4.9%)		1147 (4.2%)	1229 (4.2%)	
2015	15 (4.3%)	14 (3.8%)		2025 (7.4%)	2091 (7.2%)	
2016	37 (10.6%)	34 (9.3%)		2647 (9.6%)	2791 (9.6%)	

Table 2. *Cont.*

	ROS1-Positive			ROS1-Negative		
	Abstracted Cohort	**ML-Extracted Cohort**	**aSMD**	**Abstracted Cohort**	**ML-Extracted Cohort**	**aSMD**
2017	56 (16.0%)	52 (14.2%)		3487 (12.7%)	3719 (12.7%)	
2018	44 (12.6%)	46 (12.5%)		3726 (13.6%)	3949 (13.5%)	
2019	47 (13.5%)	44 (12.0%)		3811 (13.9%)	3997 (13.7%)	
2020	36 (10.3%)	49 (13.4%)		3713 (13.5%)	3812 (13.0%)	
2021	49 (14.0%)	53 (14.4%)		3708 (13.5%)	3917 (13.4%)	
2022	22 (6.3%)	30 (8.2%)		2228 (8.1%)	2661 (9.1%)	
Group stage, n (%)			0.10			0.06
Stage I	16 (4.6%)	17 (4.6%)		2331 (8.5%)	2582 (8.8%)	
Stage II	5 (1.4%)	5 (1.4%)		1387 (5.0%)	1438 (4.9%)	
Stage III	60 (17.2%)	53 (14.4%)		5514 (20.1%)	5832 (20.0%)	
Stage IV	262 (75.1%)	288 (78.5%)		17,692 (64.4%)	18,999 (65.0%)	
Group stage is not reported	6 (1.7%)	4 (1.1%)		554 (2.0%)	368 (1.3%)	
Histology, n (%)			0.08			0.04
Non-squamous cell carcinoma	313 (89.7%)	334 (91.0%)		20,266 (73.8%)	21,880 (74.9%)	
NSCLC histology NOS	12 (3.4%)	8 (2.2%)		1274 (4.6%)	1155 (4.0%)	
Squamous cell carcinoma	24 (6.9%)	25 (6.8%)		5938 (21.6%)	6184 (21.2%)	
ECOG PS at advanced diagnosis, n (%)			0.10			0.02
0	86 (24.6%)	82 (22.3%)		5549 (20.2%)	5985 (20.5%)	
1	99 (28.4%)	99 (27.0%)		7762 (28.2%)	8405 (28.8%)	
2	18 (5.2%)	17 (4.6%)		2588 (9.4%)	2788 (9.5%)	
3	5 (1.4%)	4 (1.1%)		618 (2.2%)	632 (2.2%)	
4	0 (0.0%)	0 (0.0%)		32 (0.1%)	34 (0.1%)	
Missing/not documented	141 (40.4%)	165 (45.0%)		10,929 (39.8%)	11,375 (38.9%)	
PD-L1 status, n (%)			0.09			0.07
Negative	57 (16.3%)	56 (15.3%)		6548 (23.8%)	6878 (23.5%)	
Positive	178 (51.0%)	189 (51.5%)		12,500 (45.5%)	13,162 (45.0%)	
Unknown	21 (6.0%)	30 (8.2%)		1117 (4.1%)	1614 (5.5%)	
Not tested	93 (26.6%)	92 (25.1%)		7313 (26.6%)	7565 (25.9%)	
Treatment received, n (%)			0.13			0.03
Non-oral antineoplastic	51 (14.6%)	68 (18.5%)		19,505 (71.0%)	20,662 (70.7%)	
Other oral therapy	36 (10.3%)	33 (9.0%)		2691 (9.8%)	2674 (9.2%)	
ROS1 inhibitor	224 (64.2%)	220 (59.9%)		159 (0.6%)	139 (0.5%)	
No treatment documented	38 (10.9%)	46 (12.5%)		5123 (18.6%)	5744 (19.7%)	

Abbreviations: aSMD: absolute standardized mean difference; ECOG PS: Eastern Cooperative Oncology Group performance status; IQR: interquartile range; NSCLC: non-small cell lung cancer; PD-L1: programmed; death-ligand 1; [a] Patients who reported Hispanic or Latinx ethnicity, regardless of race, were included in *Other race*.

3.1.2. Describing Natural History of Disease in Biomarker Sub-Groups

The natural history analysis of rwOS in biomarker sub-groups found the same conclusions using expert-abstracted data as with the replication using ML-extracted data. Both curation techniques found that lung cancer patients with a positive biomarker result for *ROS1* lived longer than patients with a negative result (Figure 3). From expert-abstracted data, the median rwOS was 11.28 months (95% CI: 11.02, 11.51) and 19.57 (95% CI: 17.34, 28.20) months for patients with a biomarker-negative and -positive test result, respectively. Replicating the analysis with ML-extracted data, the median rwOS was 11.05 months (95% CI: 10.82, 11.31) and 18.20 months (95% CI: 15.61, 22.79) for patients who were biomarker-negative and -positive, respectively. The relative association between biomarker result and survival did not differ between the expert-abstracted and ML-extracted data, where similar HRs and standard errors were observed (Table 3). Further, a robustness check, statistically adjusting for variables associated with *ROS1* result or survival, reached similar conclusions (Supplementary Table S2).

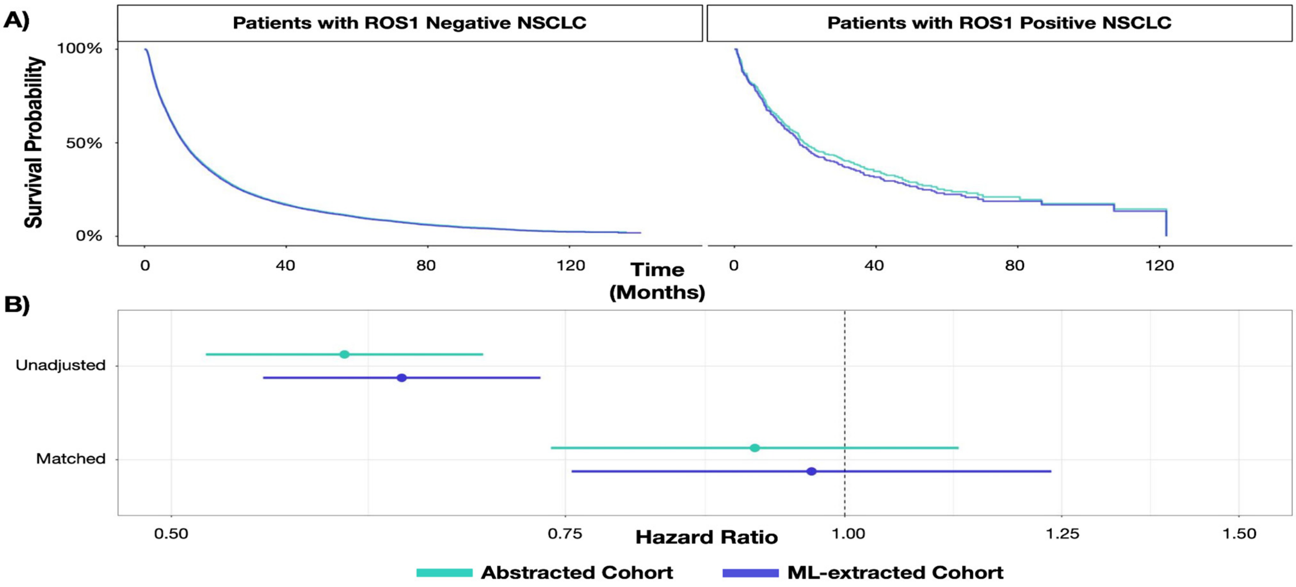

Figure 3. Results from replication of natural history study. Abbreviations: ML: machine learning; NSCLC: non-small cell lung cancer. (Panel (**A**)): Kaplan–Meier curves for patients with *ROS1*-positive and -negative NSCLC by data curation approach. (Panel (**B**)): Association between *ROS1* status and survival by data curation approach.

Table 3. Association between *ROS1* biomarker status and survival, by data curation approach.

	RWD Curation Approach	Biomarker Overall Survival HR (95% CI)	SE	*p*-Value
Unadjusted analysis	Expert-abstracted data	0.60 (0.52, 0.69)	0.073	$p < 0.001$
	ML-extracted data	0.63 (0.55, 0.73)	0.073	$p < 0.001$
Adjusted analysis	Expert-abstracted data	0.91(0.74, 1.12)	0.107	0.387
	ML-extracted data	0.97 (0.76, 1.24)	0.126	0.785

Abbreviations: CI: confidence interval; HR: hazard ratio; ML: machine learning; SE: standard error.

3.2. Treatment-Defined Cohort

Selection of the treatment-defined cohort included 682 patients when using data curated by expert abstraction and 701 patients when using data curated by ML extraction. The BCP treatment utilization rate was 34.60% (expert-abstracted data) and 34.52% (ML-extracted data) with other patients receiving the CP treatment regimen. Patient attrition for this cohort when using both expert-abstracted and ML-extracted data is described in Supplementary Table S3.

3.2.1. Balancing Populations

There was no meaningful difference in the distribution of treatment propensity score weights based on the datasets having expert-abstracted compared to ML-extracted variables. After applying inverse propensity score weights to the cohorts, we observed a similar covariate balance between treatment groups in both cohorts (Figure 4). Both weighted cohorts achieved balance (absolute or standardized mean difference < 0.1) across all variables, with the exception of the treatment start year, which has a slight residual imbalance in both the abstracted and ML-extracted cohorts.

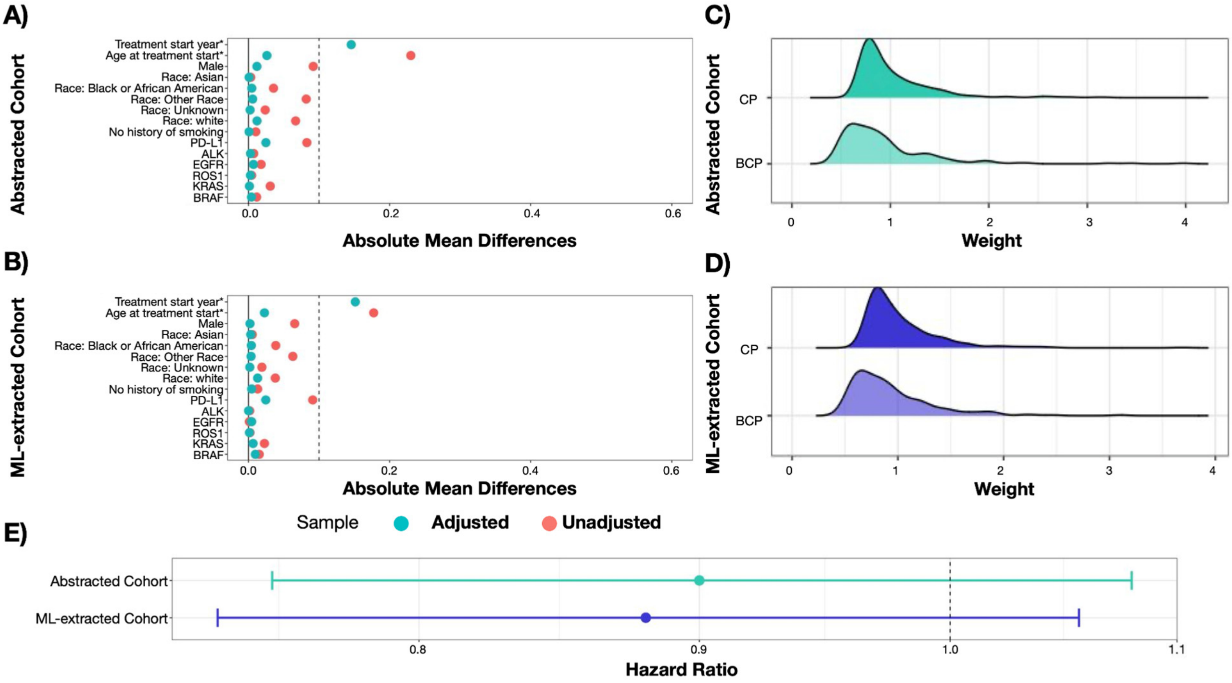

Figure 4. Results from replication of comparative effectiveness study. Abbreviations: 1L: first line; *ALK*: anaplastic lymphoma kinase; *EGFR*: epidermal growth factor receptor; PD-L1: programmed death-ligand 1; ML: machine learning. (Panel (**A**)): Covariate balance plot, abstracted cohort. (Panel (**B**)): Covariate balance plot, ML-extracted cohort. (Panel (**C**)): Distribution of weights stratified by treatment group, abstracted cohort. (Panel (**D**)): Distribution of weights stratified by treatment group, ML-extracted cohort. (Panel (**E**)): Effect of treatment group on survival, stratified by data curation approach.

3.2.2. Measuring Treatment Comparative Effectiveness

There was no meaningful difference in the result of the estimated treatment HR for rwOS based on datasets containing expert-abstracted compared to ML-extracted variables (Table 4). With expert-abstracted data, the estimated HR was 0.90 (95% CI: 0.75, 1.08), indicating slightly longer survival for patients who received BCP compared with CP. With ML-extracted data, the HR was 0.88 (95% CI: 0.74, 1.06). The HR confidence intervals were similar between expert-abstracted and ML-extracted replication, and they yielded the same statistical inference.

Table 4. Association between treatment regimen (CP vs. BCP) and survival, by data curation approach.

RWD Curation Approach	Treatment Effectiveness HR (95% CI)	SE	*p*-Value
Expert-abstracted data	0.90 (0.75, 1.08)	0.092	0.258
ML-extracted data	0.88 (0.74, 1.06)	0.092	0.170

Abbreviations: BCP: bevacizumab–carboplatin–paclitaxel; CI: confidence interval; CP: carboplatin–paclitaxel; HR: hazard ratio; ML: machine learning; SE: standard error.

4. Discussion

This study assessed the quality and fitness-for-use of oncology EHR-derived data curated with ML-extracted variables as compared to the reference standard of expert-abstracted variables.

We replicated four common observational research archetypes for EHR-derived datasets where the analytic cohort was defined first with abstracted and second with ML-extracted data. Overall, there was no meaningful statistical or clinical difference in the results based on ML-extracted variables in reference to the benchmark of expert abstraction. In a biomarker-defined patient population, we observed similar distributions of patient characteristics. Moreover, the conclusions about an association between biomarker status and survival was consistent between data curation approaches. Likewise, in the treatment-defined cohort, the distribution of the propensity score weights was similar for expert-abstracted and ML-extracted data. The replication of a treatment comparative effectiveness analysis also produced the same results. Together, these findings demonstrate that evidence generated by analyzing ML-extracted data can lead to the same conclusions as evidence generated with abstracted data when ML models are trained on expert-labeled data and evaluated with a research-centric approach.

We showed how more efficiently curated ML-extracted data can replicate the distribution of baseline patient characteristics that were alternatively generated through labor-intensive expert abstraction from charts. This opens more opportunities to study niche populations with larger cohorts as well as adjust for potential confounders in these patient populations with confidence that the data curation approach made no difference in the study findings or conclusions.

Given the design of our study, observed differences can be attributed to variability in how a patient's unstructured data were processed by abstractors and ML models, resulting in patients' observed data values being discordant. Nevertheless, minor differences in the generated evidence were observed when using expert-abstracted and ML-extracted RWD. These differences did not exceed more than an 8% difference, nor did any difference amount to what would be considered statistically or clinically meaningful.

Estimates of biomarker-associated prevalence and survival obtained using ML-extracted data are consistent with previous studies [23–27]. Additionally, the comparative treatment effect measured in the treatment-defined cohort is consistent with similar comparisons found in the literature [28] as well as with clinical trials [28,29]. While this analysis was not powered to demonstrate a difference, the consistency of the results obtained using ML-extracted data with the results using expert-abstracted data and from external studies further highlights the fact that RWE based on ML-extracted data are reliable when obtained from an adequate and well controlled study.

A side effect of data misclassification is the distortion of type I and II error rates [30,31]. While misclassification in the ML-extracted data may exist, it did not lead to meaningfully different model standard errors. Decision makers such as payers and health technology assessment (HTA) bodies can evaluate evidence generated using ML-extracted data similarly to evidence generated with expert-abstracted data. As misclassification is a limitation of observational research, researchers who use unstructured RWD in their studies, regardless of the curation method, should continue to apply quantitative bias analyses [32] or other bias correction methods [33,34] to understand the potential impact of misclassification.

While ML extraction can generate fit-for-purpose data for observational research, there are a number of challenges that represent significant hurdles to more widespread adoption. This includes the need for generalizable, high-quality, labeled data to train ML models in order to sufficiently reflect the target population and avoid a potential bias or inadvertent exclusion of historically marginalized populations [13,35]. Low quality or noisy labels may distort the learned function between features and labels, which could lead to incorrect model predictions and/or poor model performance. Additionally, there is a need for model transparency and explainability such that model predictions can be trusted by stakeholders and therefore be more readily accepted [36]. Finally, proper model evaluation is needed to ensure that models are fair and generalizable, which requires an adequate volume of high-quality labeled test data that is not used during model training and validation [6,37].

The findings of this study should be viewed considering certain limitations. First, this study demonstrates the fit-for-purpose of an ML-extracted dataset using a limited number of results spanning two analytic cohorts. It is possible that for another study population of interest, there could be differences between the results obtained using abstracted data vs. ML-extracted data. Nevertheless, the ML-extracted variables used in this analysis were trained on high-volume, high-quality abstracted data from a large nationwide database. Additional analyses to demonstrate that ML-extracted data are fit-for-purpose and can unlock new use cases are planned for different patient populations; however, given our sample sizes and use of expert-abstracted training data, we believe we will obtain similar results. Second, this study was not implemented on a dataset independent of model development. To do so would require abstracting an additional 186,000 lung cancer patients to obtain similar cohort sizes observed in the presented analyses. While the dataset used here is not independent of model development, it is important to note that the tasks that the models were trained to perform (i.e., information extraction) are independent of the analyses performed in this study. Third, although we adjusted for potential confounders, including demographics and relevant clinical factors, there is potential bias from confounding by unmeasured covariates, missingness, treatment compliance, or measurement error. However, it is important to note that these sources of biases will similarly impact the results regardless of the data curation approach; therefore, the comparison is unlikely to be impacted. Finally, while the ML-extracted dataset used in this study draws from multiple cancer centers that are representative of patients with cancer in the US, [15] this study does not evaluate the generalizability of these models to external cancer centers that were not included in the training population. Although the models themselves are not necessarily transportable and would benefit from retraining before use in other populations, [38] this study demonstrates that the evidence generated from well-designed pharmacoepidemiological studies using a representative cohort with ML-enabled clinical depth can be generalizable.

5. Conclusions

In our study, we assessed the reproducibility of oncology RWE studies using ML-extracted variables in reference to the benchmark of the standard approach in retrospective research studies with manual chart review. We performed multiple example analyses representing common archetypes for the application of EHR data in oncology research and evaluated their results in support of developing reliable, fit-for-purpose RWD using ML extraction. Our results showed that ML-extracted variables can produce similar results and analytic conclusions of analyses based on expert-abstracted variables. The ability to extract high-quality data at scale through ML extraction has the potential to unlock valuable insights and advance clinical and health outcomes research, especially when quality is more broadly communicated and understood.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/cancers15061853/s1, Table S1: Biomarker-defined cohort attrition table; Table S2: Robustness check of the association between ROS1 status and survival, by data curation approach; Table S3: Treatment-defined cohort attrition table. References [39–41] have been cited in Supplementary Materials.

Author Contributions: Conceptualization, C.M.B., A.B.C., S.N., M.E. and S.B.; methodology, C.M.B., A.S., E.F., A.B.C., M.E. and S.B.; validation, C.M.B. and A.S.; formal analysis, C.M.B. and A.S.; writing—original draft preparation, C.M.B., A.S., E.F., B.A. and M.E.; writing—review and editing, C.M.B., A.S., E.F., A.B.C., S.N., B.A., M.E. and S.B.; supervision, A.B.C., M.E. and S.B.; project administration, S.N., M.E. and S.B. All authors have read and agreed to the published version of the manuscript.

Funding: This study was sponsored by Flatiron Health, Inc., which is an independent subsidiary of the Roche Group.

Institutional Review Board Statement: Institutional Review Board approval of the study protocol was obtained prior to study conduct and included a waiver of informed consent.

Informed Consent Statement: Patient consent was waived due to the following; does not involve greater than minimal risk; leverages observational research, which relies on data which was previously collected—as such it is not practicable to conduct the research without the waiver or alteration; and waiving or altering the informed consent will not adversely affect the subjects' rights and welfare.

Data Availability Statement: The data that support the findings of this study have been originated by Flatiron Health, Inc. Requests for data sharing by license or by permission for the specific purpose of replicating results in this manuscript can be submitted to dataaccess@flatiron.com.

Acknowledgments: The authors would like to thank Hannah Gilham of Flatiron Health for editorial and project management support, as well as Flatiron Health employees Guy Amster, John Ritten, Samuel Wilkinson, Jonathan Kelly, Konstantin Krismer, James Gippetti, Auriane Blarre, Michael Waskom, Katherine Harrison, George Ho, Robin Linzmayer, and Chaya Wurman.

Conflicts of Interest: At the time of the study, all authors report employment at Flatiron Health, Inc., an independent subsidiary of the Roche Group, and stock ownership in Roche. ME and AC report equity ownership in Flatiron Health, Inc. (initiated before acquisition by Roche in April 2018).

References

1. Guinn, D.; Wilhelm, E.E.; Lieberman, G.; Khozin, S. Assessing function of electronic health records for real-world data generation. *BMJ Evid.-Based Med.* **2019**, *24*, 95–98. [CrossRef]
2. Stark, P. Congressional intent for the HITECH Act. *Am. J. Manag. Care* **2010**, *16*, SP24–SP28. Available online: https://www.ncbi.nlm.nih.gov/pubmed/21314216 (accessed on 12 January 2023).
3. Stewart, M.; Norden, A.D.; Dreyer, N.; Henk, H.J.; Abernethy, A.P.; Chrischilles, E.; Kushi, L.; Mansfield, A.S.; Khozin, S.; Sharon, E.; et al. An Exploratory Analysis of Real-World End Points for Assessing Outcomes Among Immunotherapy-Treated Patients with Advanced Non–Small-Cell Lung Cancer. *JCO Clin. Cancer Inform.* **2019**, *3*, 1–15. [CrossRef]
4. Zhang, J.; Symons, J.; Agapow, P.; Teo, J.T.; Paxton, C.A.; Abdi, J.; Mattie, H.; Davie, C.; Torres, A.Z.; Folarin, A.; et al. Best practices in the real-world data life cycle. *PLoS Digit. Health* **2022**, *1*, e0000003. [CrossRef]
5. Birnbaum, B.; Nussbaum, N.; Seidl-Rathkopf, K.; Agrawal, M.; Estevez, M.; Estola, E.; Haimson, J.; He, L.; Larson, P.; Richardson, P. Model-assisted cohort selection with bias analysis for generating large-scale cohorts from the EHR for oncology research. *arXiv* **2020**, arXiv:2001.09765. [CrossRef]
6. Estevez, M.; Benedum, C.M.; Jiang, C.; Cohen, A.B.; Phadke, S.; Sarkar, S.; Bozkurt, S. Considerations for the Use of Machine Learning Extracted Real-World Data to Support Evidence Generation: A Research-Centric Evaluation Framework. *Cancers* **2022**, *14*, 3063. [CrossRef]
7. Koskimaki, J.; Hu, J.; Zhang, Y.; Mena, J.; Jones, N.; Lipschultz, E.; Vaidya, V.P.; Altay, G.; Erese, V.A.; Swaminathan, K.K.; et al. Natural language processing-optimized case selection for real-world evidence studies. *JCO* **2022**, *40*, 1556. [CrossRef]
8. Padula, W.V.; Kreif, N.; Vanness, D.J.; Adamson, B.; Rueda, J.-D.; Felizzi, F.; Jonsson, P.; Ijzerman, M.J.; Butte, A.; Crown, W. Machine Learning Methods in Health Economics and Outcomes Research—The PALISADE Checklist: A Good Practices Report of an ISPOR Task Force. *Value Health* **2022**, *25*, 1063–1080. [CrossRef]
9. US Food and Drug Administration. Good Machine Learning Practice for Medical Device Development: Guiding Principles. 2021. Available online: https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles.http://elsibi.hypotheses.org/3154 (accessed on 2 November 2022).
10. NICE Real-World Evidence Framework. National Institute for Health and Care Excellence Web Site. Available online: https://www.nice.org.uk/corporate/ecd9/chapter/overview (accessed on 27 February 2023).
11. Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products; Draft Guidance for Industry; Availability. U.S. Food & Drug Administration Documents/FIND. 2021. Available online: https://www.fda.gov/media/152503/download (accessed on 2 November 2022).
12. Schurman, B. The Framework for FDA's Real-World Evidence Program. *Appl. Clin. Trials* **2019**, *28*, 15–17. Available online: https://search.proquest.com/docview/2228576959 (accessed on 2 November 2022).
13. Obermeyer, Z.; Powers, B.; Vogeli, C.; Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **2019**, *366*, 447–453. [CrossRef]
14. Maarseveen, T.D.; Maurits, M.P.; Niemantsverdriet, E.; Mil, A.H.M.V.D.H.-V.; Huizinga, T.W.J.; Knevel, R. Handwork vs machine: A comparison of rheumatoid arthritis patient populations as identified from EHR free-text by diagnosis extraction through machine-learning or traditional criteria-based chart review. *Arthritis Res. Ther.* **2021**, *23*, 174. [CrossRef]
15. Ma, X.; Long, L.; Moon, S.; Adamson, B.J.S.; Baxi, S.S. Comparison of Population Characteristics in Real-World Clinical Oncology Databases in the US: Flatiron Health, SEER, and NPCR. *medRxiv* **2020**. [CrossRef]
16. Zhang, Q.; Gossai, A.; Monroe, S.; Nussbaum, N.C.; Parrinello, C.M. Validation analysis of a composite real-world mortality endpoint for patients with cancer in the United States. *Health Serv. Res.* **2021**, *56*, 1281–1287. [CrossRef]

17. Adamson, B.J.; Waskom, M.; Blarre, A.; Kelly, J.; Krismer, K.; Nemeth, S.; Gippetti, J.; Ritten, J.; Harrison, K.; Ho, G.; et al. Approach to Machine Learning for Extraction of Real-World Data Variables from Electronic Health Records. *medRxiv* **2023**. [CrossRef]
18. Austin, P.C. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivar. Behav. Res.* **2011**, *46*, 399–424. [CrossRef]
19. Tsai, W.-Y.; Jewell, N.P.; Wang, M.-C. A note on the product-limit estimator under right censoring and left truncation. *Biometrika* **1987**, *74*, 883–886. [CrossRef]
20. Sondhi, A. Estimating survival parameters under conditionally independent left truncation. *Pharm. Stat.* **2022**, *21*, 895–906. [CrossRef]
21. Zhou, C.; Wu, Y.-L.; Chen, G.; Liu, X.; Zhu, Y.; Lu, S.; Feng, J.; He, J.; Han, B.; Wang, J.; et al. BEYOND: A Randomized, Double-Blind, Placebo-Controlled, Multicenter, Phase III Study of First-Line Carboplatin/Paclitaxel Plus Bevacizumab or Placebo in Chinese Patients With Advanced or Recurrent Nonsquamous Non–Small-Cell Lung Cancer. *JCO* **2015**, *33*, 2197–2204. [CrossRef]
22. R Core Team. R: A Language and Environment for Statistical Computing. Available online: https://www.r-project.org/ (accessed on 2 November 2022).
23. Doebele, R.C.; Perez, L.; Trinh, H.; Martinec, M.; Martina, R.; Riehl, T.; Krebs, M.G.; Meropol, N.J.; Wong, W.B.; Crane, G. Comparative effectiveness analysis between entrectinib clinical trial and crizotinib real-world data in *ROS1* + NSCLC. *J. Comp. Eff. Res.* **2021**, *10*, 1271–1282. [CrossRef]
24. Ahmadzada, T.; Kao, S.; Reid, G.; Boyer, M.; Mahar, A.; Cooper, W.A. An Update on Predictive Biomarkers for Treatment Selection in Non-Small Cell Lung Cancer. *J. Clin. Med.* **2018**, *7*, 153. [CrossRef]
25. Gadgeel, S.M.; Thakur, M.K. Predictive and Prognostic Biomarkers in Non-Small Cell Lung Cancer. *Semin. Respir. Crit. Care Med.* **2016**, *37*, 760. [CrossRef]
26. Thunnissen, E.; Van Der Oord, K.; Bakker, M.D. Prognostic and predictive biomarkers in lung cancer. A review. *Virchows Arch.* **2014**, *464*, 347–358. [CrossRef]
27. Tu, H.; Wu, M.; Huang, W.; Wang, L. Screening of potential biomarkers and their predictive value in early stage non-small cell lung cancer: A bioinformatics analysis. *Transl. Lung Cancer Res.* **2019**, *8*, 797–807. [CrossRef] [PubMed]
28. Liu, Y.; Li, H.-M.; Wang, R. Effectiveness and Safety of Adding Bevacizumab to Platinum-Based Chemotherapy as First-Line Treatment for Advanced Non-Small-Cell Lung Cancer: A Meta-Analysis. *Front. Med.* **2021**, *8*, 616380. [CrossRef]
29. Sandler, A.; Gray, R.; Perry, M.C.; Brahmer, J.; Schiller, J.H.; Dowlati, A.; Lilenbaum, R.; Johnson, D.H. Paclitaxel–Carboplatin Alone or with Bevacizumab for Non–Small-Cell Lung Cancer. *N. Engl. J. Med.* **2006**, *355*, 2542–2550. [CrossRef] [PubMed]
30. Chen, Y.; Wang, J.; Chubak, J.; Hubbard, R.A. Inflation of type I error rates due to differential misclassification in EHR-derived outcomes: Empirical illustration using breast cancer recurrence. *Pharmacoepidemiol. Drug Saf.* **2019**, *28*, 264–268. [CrossRef]
31. Van Smeden, M.; Lash, T.L.; Groenwold, R.H.H. Reflection on modern methods: Five myths about measurement error in epidemiological research. *Int. J. Epidemiol.* **2020**, *49*, 338–347. [CrossRef]
32. Lash, T.L.; Fox, M.P.; MacLehose, R.F.; Maldonado, G.; McCandless, L.C.; Greenland, S. Good practices for quantitative bias analysis. *Int. J. Epidemiol.* **2014**, *43*, 1969–1985. [CrossRef] [PubMed]
33. Wang, S.; McCormick, T.H.; Leek, J.T. Methods for correcting inference based on outcomes predicted by machine learning. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 30266–30275. [CrossRef]
34. Richardson, S.; Gilks, W.R. A Bayesian Approach to Measurement Error Problems in Epidemiology Using Conditional Independence Models. *Am. J. Epidemiol.* **1993**, *138*, 430–442. [CrossRef]
35. Berger, M.L.; Curtis, M.D.; Smith, G.; Harnett, J.; Abernethy, A.P. Opportunities and challenges in leveraging electronic health record data in oncology. *Futur. Oncol.* **2016**, *12*, 1261–1274. [CrossRef]
36. Tayefi, M.; Ngo, P.; Chomutare, T.; Dalianis, H.; Salvi, E.; Budrionis, A.; Godtliebsen, F. Challenges and opportunities beyond structured data in analysis of electronic health records. *WIREs Comput. Stat.* **2021**, *13*, e1549. [CrossRef]
37. Hernandez-Boussard, T.; Bozkurt, S.; Ioannidis, J.P.A.; Shah, N.H. MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care. *J. Am. Med. Inform. Assoc.* **2020**, *27*, 2011–2015. [CrossRef] [PubMed]
38. Coquet, J.; Bievre, N.; Billaut, V.; Seneviratne, M.; Magnani, C.J.; Bozkurt, S.; Brooks, J.D.; Hernandez-Boussard, T. Assessment of a Clinical Trial–Derived Survival Model in Patients With Metastatic Castration-Resistant Prostate Cancer. *JAMA Netw. Open* **2021**, *4*, e2031730. [CrossRef] [PubMed]
39. Mansournia, M.A.; Hernán, M.A.; Greenland, S. Matched Designs and Causal Diagrams. *Int. J. Epidemiol.* **2013**, *42*, 860–869. [CrossRef] [PubMed]
40. Stuart, E.A. Matching Methods for Causal Inference: A Review and a Look Forward. *Stat. Sci.* **2010**, *25*, 1–21. [CrossRef] [PubMed]
41. Vander Weele, T.J.; Shpitser, I. A New Criterion for Confounder Selection. *Biometrics* **2011**, *67*, 1406–1413. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.