

Article

A Natural Language Processing Algorithm to Improve Completeness of ECOG Performance Status in Real-World Data

Aaron B. Cohen ^{1,2,*} , Andrej Rosic ¹, Katherine Harrison ¹, Madeline Richey ¹ , Sheila Nemeth ¹, Geetu Ambwani ¹, Rebecca Miksad ¹, Benjamin Haaland ³ and Chengsheng Jiang ¹

¹ Flatiron Health Inc., 233 Spring St., New York, NY 10013, USA

² Department of Medicine, NYU Grossman School of Medicine, New York, NY 10016, USA

³ Huntsman Cancer Institute, University of Utah, Salt Lake City, UT 84112, USA

* Correspondence: acohen@flatiron.com

Featured Application: Critical clinical variables, such as ECOG performance status, are required for retrospective research but may be incomplete. A natural language processing algorithm was used to improve the completeness of ECOG performance status across an electronic health record-derived database.

Abstract: Our goal was to develop and characterize a Natural Language Processing (NLP) algorithm to extract Eastern Cooperative Oncology Group Performance Status (ECOG PS) from unstructured electronic health record (EHR) sources to enhance observational datasets. By scanning unstructured EHR-derived documents from a real-world database, the NLP algorithm assigned ECOG PS scores to patients diagnosed with one of 21 cancer types who lacked structured ECOG PS numerical scores, anchored to the initiation of treatment lines. Manually abstracted ECOG PS scores were used as a source of truth to both develop the algorithm and evaluate accuracy, sensitivity, and positive predictive value (PPV). Algorithm performance was further characterized by investigating the prognostic value of composite ECOG PS scores in patients with advanced non-small cell lung cancer receiving first line treatment. Of N = 480,825 patient-lines, structured ECOG PS scores were available for 290,343 (60.4%). After applying NLP-extraction, the availability increased to 73.2%. The algorithm's overall accuracy, sensitivity, and PPV were 93% (95% CI: 92–94%), 88% (95% CI: 87–89%), and 88% (95% CI: 87–89%), respectively across all cancer types. In a cohort of N = 51,948 aNSCLC patients receiving 1L therapy, the algorithm improved ECOG PS completeness from 61.5% to 75.6%. Stratification by ECOG PS showed worse real-world overall survival (rWOS) for patients with worse ECOG PS scores. We developed an NLP algorithm to extract ECOG PS scores from unstructured EHR documents with high accuracy, improving data completeness for EHR-derived oncology cohorts.

Keywords: EHR; machine learning; ECOG PS; RWD; RWE; NLP; cancer diagnosis; natural language processing



Citation: Cohen, A.B.; Rosic, A.; Harrison, K.; Richey, M.; Nemeth, S.; Ambwani, G.; Miksad, R.; Haaland, B.; Jiang, C. A Natural Language Processing Algorithm to Improve Completeness of ECOG Performance Status in Real-World Data. *Appl. Sci.* **2023**, *13*, 6209. <https://doi.org/10.3390/app13106209>

Academic Editor: José Machado

Received: 7 March 2023

Revised: 7 April 2023

Accepted: 15 May 2023

Published: 18 May 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Real-world data (RWD), defined as clinical data collected in the course of routine medical care, and real-world evidence (RWE), the insights gathered via analysis and interpretation of those data [1], have become major components of the clinical research landscape. RWD rely on several possible sources, such as administrative claims, registries or electronic health records (EHRs). The adoption of new technologies in the clinic, the digitization of clinical information management, and analytical advances in particular have unlocked the potential for the use of EHRs as a major RWD source for clinical research [2–6].

However, EHRs have known limitations as research data sources. Information collected during routine patient care within EHRs varies between practices and may be driven by different EHR software or clinical workflows. Critical data points related to oncology

care may not be comprehensively captured in structured fields and rather captured in an unstructured format [7]. In addition, the dynamics of routine care affect documentation practices. For example, documentation for purposes of patient care may be less complete because data elements pertinent for supporting decision making or billing may be prioritized; thus, normal values may be less likely to be captured compared to abnormal. As a result, incomplete or missing data and documentation variability may be suboptimal in the use of EHR-derived data for research [8–11]. One approach to optimize the completeness of EHR-derived data is to combine, via manual abstraction, unstructured information (e.g., from physician notes and scanned documents) with structured data captured in predefined EHR data fields [2,12,13]. This approach can unlock information found only in unstructured formats and also leverage potential overlaps between unstructured and structured sources. For instance, treating physicians may describe general health status using predefined EHR fields and document it in their notes. Consulting both these sources may unveil nuances in clinical information and help resolve discrepancies during quality control processes. However, manual abstraction requires curation, detailed procedures and policies, and careful quality controls; therefore, overall scalability may be limited by its resource intensiveness [2,13,14].

Natural language processing (NLP) has emerged as a valuable approach to automate information extraction from unstructured clinical data sources, becoming a fertile ground for the development of tools in the healthcare setting [15–21]. NLP applications to support clinical decision making range from interpretation of diagnostic tests [22,23] to the stratification of patients based on unstructured EHR data [24–27]. Within the field of oncology, NLP can also be deployed to improve the quality and completeness of important clinical details, including stage, recurrence, and labs [28–33].

Eastern Cooperative Oncology Group performance status (ECOG PS) score is a critical variable in oncology retrospective research. ECOG PS is a qualitative numeric score indicative of the general health status of a patient with cancer. The score ranges from 0 (no limitations) to 5 (deceased) and has been consistently shown to have strong prognostic value [34]. ECOG PS score is commonly evaluated by clinicians, particularly to inform management decisions, such as whether to initiate a new treatment. ECOG PS is also a standard clinical trial eligibility criterion, a key stratification factor, and an important variable to include in multivariable analyses in oncology studies [35–40]. However, depending on EHR work-flows and physician preference, ECOG PS may not always be recorded as a structured data point, or even documented at all [34,41–43].

In our study, we developed and evaluated an NLP algorithm to extract ECOG PS scores from unstructured EHR sources at the time of new treatment initiation, when ECOG PS most impacts clinical decision making. NLP extraction occurred across 21 distinct cancer types and supplemented available structured scores to enhance data completeness. We used a nationwide EHR-derived RWD cohort with manually abstracted ECOG PS scores as a reference source. We further investigated the prognostic value of composite (structured and NLP-extracted) ECOG PS scores in a cohort of patients with advanced NSCLC (aNSCLC) receiving first line (1L) treatment by comparing real-world overall survival (rwOS) for patients with structured vs. extracted ECOG PS scores.

2. Related Work

Prior efforts to extract ECOG PS score using NLP exist in the literature, but our research has some key distinctions [44–48]. First our algorithm was developed and tested on a nationwide dataset with approximately 280 US cancer clinics spanning both community and academic settings. In contrast, past ECOG PS NLP efforts have been developed using smaller cohorts primarily from single sites of cancer care [44,45]. Additionally, previous work has only been carried out within a single disease, or indexed around time of diagnosis, but this study developed an algorithm that works across 21 cancer types and extracts ECOG PS scores around treatment initiation across a patient’s care journey [45–47]. Our algorithm also focused on supplementing structured ECOG PS scores with NLP-extracted ECOG PS

scores from unstructured data to create a composite ECOG PS score that provides a more complete picture of a patient’s ECOG performance status from RWD [47,48]. To further ensure fit-for-use with RWE, this study was unique in evaluating the composite score for prognostic use as well as stratifying performance by ECOG PS score [44,46].

3. Materials and Methods

3.1. Data Source

This study used the nationwide electronic health record (EHR)-derived de-identified Flatiron Health database. This is a longitudinal database, comprising de-identified patient-level structured and unstructured data [49], curated via technology-enabled abstraction [13]. During the study period, the de-identified data originated from approximately 280 US cancer clinics (~800 sites of care). Institutional Review Board approval of the study protocol was obtained prior to study conduct, and included a waiver of informed consent.

This study used EHR-derived de-identified data for patients diagnosed after 2011 with at least one of the 21 cancer types: acute myeloid leukemia (AML), metastatic breast cancer (mBC), chronic lymphocytic leukemia (CLL), metastatic colorectal cancer (mCRC), diffuse large B-cell lymphoma (DLBCL), early breast cancer (eBC), endometrial cancer, follicular lymphoma (FL), advanced gastro-esophageal cancer (aGE), hepatocellular carcinoma (HCC), advanced head and neck cancer (aHNC), mantle cell lymphoma (MCL), advanced melanoma (aMel), multiple myeloma (MM), aNSCLC, ovarian cancer, metastatic pancreatic cancer, metastatic prostate cancer, metastatic renal-cell carcinoma (mRCC), small cell lung cancer (SCLC), and advanced urothelial cancer (detailed eligibility in the Supplement).

3.2. Components of the NLP-Extracted ECOG PS Variable

A regular-expression-based NLP algorithm was developed to scan through unstructured EHR documents (including oncology clinic visit notes, nursing notes, radiology reports, pathology reports, and other uncategorized documents) and extract ECOG PS score values, numeric (e.g., ‘0’) or non-numeric (e.g., ‘zero’ or ‘PS0’), within three words of a set of signifiers (‘ECOG’, ‘ECOG PS’, ‘Performance’, or ‘Performance Status’) (Figure 1).

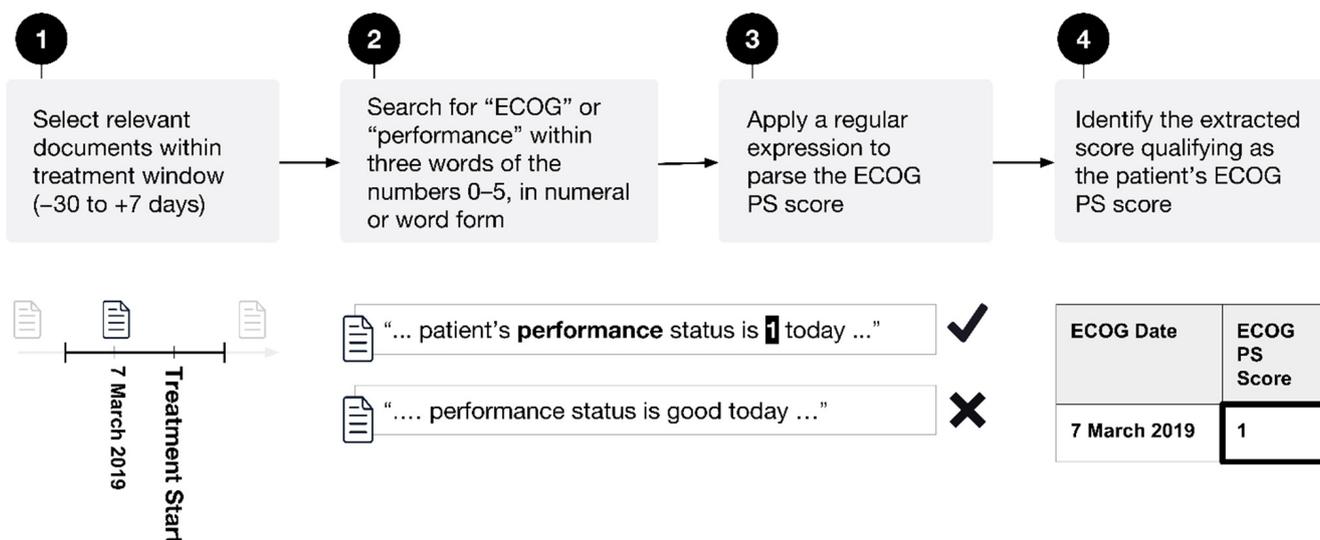


Figure 1. NLP algorithm approach to the extraction of ECOG PS scores from unstructured data sources. 1. Select relevant documents within treatment window (–30 to +7 days); 2. Search for terms such as “ECOG” or “performance” within three words of the numbers 0–5, in numeral or word form; 3. Apply a regular expression to parse the ECOG PS score; 4. Identify the extracted score qualifying as the patient’s ECOG PS score.

As inputs, the extraction algorithm receives all eligible documents in a patient’s treatment window (described below), and the regular expression is run across the full

text contents of each document. Each extracted ECOG PS score is tethered to the date of the document from which it was extracted. Through a process of manual and automated review, the base regular expression was adjusted to incorporate observed Optical Character Recognition (OCR) errors and common, alternative clinical documentation patterns not already captured, as well as made robust to erroneous ECOG-like documentation patterns not related to performance status or indicative of boiler plate text. We iteratively developed several optimizations to ensure that any form of a number (e.g., both underlined and not underlined numeric values) is picked up by the algorithm, but that numbers or dates that start with ECOG PS-like values are not mistakenly assigned as ECOG PS values [50]. The regular expression and heuristics for safe ECOG extraction were tuned across multiple held-out patient sets, with no overlap with the testing cohorts on which the approach was subsequently validated.

The categorizations by the algorithm classify patients into the following strata: PS0, PS1, PS2, PS3, and unknown. PS4 is not categorized but grouped into the unknown category due to the rare prevalence and expected resulting low performance of the algorithm. PS5 (patient deceased) is not categorized given mortality is rarely documented this way and such patients are not eligible for treatment. We anchored the extraction of ECOG PS scores to the initiation of a treatment line (period between 30 days prior through 7 days after) for a given patient. We considered this time window most clinically relevant, as ECOG PS is used to inform clinical intervention decisions. Furthermore, we allowed for ECOG PS scores up to 7 days after treatment initiation to improve completeness while taking into account lags in documentation ingestion around the time of treatment initiation. When multiple ECOG PS scores were extracted during a relevant treatment initiation window, the one closest to the treatment line index date was selected (in cases needing tie-break, scores from pre-initiation dates were selected). Demographic descriptions were generated using distinct patient, cancer-type combinations and do not reflect the number of treatment lines.

3.3. Study Cohorts

For inclusion in the study, patients needed to have been diagnosed with at least one of the 21 cancer types and have at least one treatment line defined as an anti-neoplastic therapy following the disease cohort inclusion date (detailed eligibility in the Supplement). To obtain labeled data for the training, validation, and test sets, manual abstraction of ECOG PS was performed on patients without available structured ECOG PS score in the treatment window.

A training set (consisting of 700 patients; 900 patient-lines) was used to develop the rule-based algorithm. A subsequent validation set (consisting of 1600 patients; ~2400 patient-lines) was then used for error analysis to make improvements to the initial extraction approach.

The resulting rule-based algorithm was then applied to a holdout testing set to evaluate the algorithm's performance. The testing cohort was sampled at random with no overlap with the training/validation datasets and consisted of ~7700 patient-lines (~5100 patients) without structured ECOG PS documentation in the time window before initiation of a given treatment line.

3.4. Performance Analyses

For the training, validation, and testing cohorts, the NLP algorithm was applied to patients lacking structured ECOG PS scores in the extraction time window and compared to manually-abstracted ECOG PS scores for the equivalent time window. We evaluated algorithm performance for individual ECOG PS strata as well as for binary ECOG PS categories (e.g., ECOG 0–1 vs. ECOG > 1, a cutoff commonly used for stratified analyses in clinical research).

We evaluated performance via the following metrics calculated in the testing cohort: overall accuracy (the number of assignments matching between the algorithm and the manually-abstracted information, relative to the total number of possible assignments);

sensitivity (the number of manually-abstracted known ECOG PS scores for which the algorithm identifies the same score, relative to the total number of manually-abstracted known ECOG PS scores); PPV (the number of model-extracted known ECOG PS scores that match the manually-abstracted score, relative to the total number of model-extracted known ECOG PS scores); and F1-score (the harmonized mean of sensitivity and PPV). Accuracy, sensitivity, PPV, and F1-score were analyzed on a held-out test cohort of 7709 treatment lines from 5341 patients present in the database as of July 2021. These performance metrics were also evaluated by specific strata (e.g., ECOG 0–3, ECOG < 2).

The performance of the NLP algorithm was further characterized by applying it to a cohort of patients with aNSCLC who were missing a structured ECOG PS score in their 1L treatment. We quantified the proportion of patients for whom the algorithm extracted an ECOG PS score, shifting their standing from PS-unknown to PS-known. We then investigated the prognostic value of the composite ECOG PS scores that resulted from combining patient-lines with structured scores with patient-lines with extracted scores. We first compared rwOS via KM estimates by the source of the ECOG PS score (extracted vs. structured), stratified by ECOG PS Score (0, 1, 2, 3). We additionally compared median rwOS by data source (extracted vs. structured), stratified by ECOG PS Score (0, 1, 2, 3).

4. Results

4.1. Overall Study Population and Impact on ECOG PS Completeness

As of 30 June 2021, in the overall database across all tumor types (N = 480,825 patient lines, N = 229,257 patients), structured ECOG PS scores were available for 290,343 patient lines (60.4%). After applying NLP-extraction, the availability of ECOG PS scores increased to 73.2% of patient lines. For certain treatment lines, the improvement was greater than 20% (Table S2). For example, for AML patients at 3L, our baseline completeness for ECOG PS was 43.4%, and after adding NLP ECOG PS, our completeness increased to 63.9%. Additionally, across many diseases, completeness of ECOG PS improved to a greater extent in later lines. For example, in the metastatic breast cohort, the baseline completeness of ECOG PS at first line therapy was 65.5% compared to 78.7% in fourth line (Table S2). This is perhaps attributable to clinicians being more likely to explicitly document ECOG PS in patients who have progressed through multiple treatments and are therefore likely to be sicker.

4.2. Algorithm Performance in Training and Testing Cohorts

Table 1 shows the demographic distribution of patients in the training set and testing cohorts. Across all cancer types, the NLP algorithm's overall accuracy, sensitivity, PPV, and F1-score in the testing set were 93.0% (95% CI: 92–94%), 88.0% (95% CI: 87–89%), 88.0% (95% CI: 87–89%), and 88.0% (95% CI: 87–89%), respectively (Table 2). Performance for the testing cohort was better than in the training cohort, as might be expected in this case, since the training cohort was purposefully enriched in less frequent ECOG PS strata and patients from originating care sites of small size, where algorithm performance was slightly worse.

4.3. Analysis of an aNSCLC Cohort: Impact on Sample Availability and Prognostic Value

Using a cohort of patients with aNSCLC in the 1L treatment window (N = 51,948, of whom 31,949 had structured ECOG PS scores, for 61.5% completeness), the NLP algorithm was applied to patients missing structured ECOG PS resulting in extracted ECOG PS scores for 7335 of them, increasing completeness to 75.6%. Stratifying the cohort with both structured ECOG PS (when available) and/or extracted ECOG PS (when structured ECOG PS was unavailable), rwOS analyses showed that for both structured and extracted scores, median OS worsened (18.8 to 4.8 months for patients with NLP-extracted ECOG PS and 18.4 to 4.1 months for patients with structured ECOG PS) (Table S3) as ECOG PS worsened, consistent with clinical expectations (Figure 2).

Table 1. Characteristics of patients in the study cohorts (training and testing cohorts).

Characteristic		Testing N = 5341 Unique Patient-Disease	Training N = 2519 Unique Patient-Disease	p-Value ¹
Age at 1L	18–64	2232 (41.8%)	1028 (40.8%)	>0.05
	65–74	1732 (32.4%)	781 (31.0%)	
	75 and older	1377 (25.8%)	710 (28.2%)	
Race	Asian	112 (2.1%)	45 (1.8%)	0.004
	Black or African American	469 (8.8%)	222 (8.8%)	
	Other Race	619 (11.6%)	297 (11.8%)	
	Unknown	490 (9.2%)	300 (11.9%)	
	White	3651 (68.5%)	1655 (65.7%)	
Ethnicity	Hispanic or Latino	332 (6.2%)	203 (8.0%)	0.002
	Unknown/Non-Hispanic	5009 (93.8%)	2316 (92.0%)	
Gender	F	2769 (51.9%)	1307 (51.9%)	>0.9
	M	2571 (48.1%)	1212 (48.1%)	
	(Missing)	1	0	
Practice Type	Academic	1186 (22.2%)	221 (8.8%)	<0.001
	Community	4155 (77.8%)	2298 (91.2%)	
Year of Initial/Adv/Met Diagnosis/First Treatment ²	<2018	4323 (80.9%)	2151 (85.34%)	<0.001
	≥2018	1018 (19.1%)	368 (14.6%)	
Group Stage (if applicable)	0	1 (<0.1%)	1 (<0.1%)	Not Applicable ³
	I	284 (5.3%)	114 (4.5%)	
	II	387 (7.2%)	185 (7.3%)	
	III	815 (15.3%)	405 (16.1%)	
	IV	2090 (39.1%)	1066 (42.3%)	
	Not Applicable	1764 (33.0%)	748 (29.7%)	
Year of start of 1L	<2018	3847 (72.0%)	2008 (79.7%)	<0.001
	≥2018	1494 (27.3%)	511 (20.3%)	

¹ Pearson's Chi-squared test; ² The Year of treatment is selected based on the availability order of: initial diagnosis, advanced diagnosis, metastatic diagnosis and first treatment. ³ Due to the small size of one group, the p-value is not applicable.

Table 2. Performance results of the algorithm in the training (3353 treatment lines) and testing cohorts (7709 treatment lines).

Cohort		Accuracy ^a	Sensitivity	PPV	F1-Score
ECOG 0–4 in Testing set		0.93 (0.92–0.94)	0.88 (0.87–0.89)	0.88 (0.87–0.89)	0.88 (0.87–0.89)
ECOG 0–4 in Training set ^b		0.83 (0.82–0.84)	0.80 (0.78–0.82)	0.75 (0.73–0.77)	0.77 (0.76–0.78)
Testing Cohort	ECOG PS 0	0.98 (0.98–0.98)	0.90 (0.88–0.92)	0.89 (0.87–0.91)	0.90 (0.89–0.91)
	ECOG PS 1	0.96 (0.96–0.96)	0.88 (0.86–0.90)	0.88 (0.86–0.90)	0.88 (0.87–0.89)
	ECOG PS 2	0.98 (0.98–0.98)	0.85 (0.81–0.89)	0.84 (0.80–0.88)	0.84 (0.83–0.85)
	ECOG PS 3	1.00 (1.00–1.00)	0.75 (0.67–0.83)	0.89 (0.83–0.95)	0.81 (0.80–0.82)
	ECOG PS 0–1	0.95 (0.95–0.95)	0.91 (0.90–0.92)	0.91 (0.90–0.92)	0.91 (0.90–0.92)
	ECOG PS 2–4	0.98 (0.98–0.98)	0.84 (0.81–0.87)	0.85 (0.82–0.88)	0.85 (0.84–0.86)
Training Cohort	ECOG PS 0	0.95 (0.94–0.96)	0.74 (0.70–0.78)	0.84 (0.80–0.88)	0.79 (0.78–0.80)
	ECOG PS 1	0.94 (0.93–0.95)	0.78 (0.75–0.81)	0.87 (0.84–0.90)	0.83 (0.82–0.84)
	ECOG PS 2	0.98 (0.98–0.98)	0.81 (0.76–0.86)	0.88 (0.83–0.93)	0.84 (0.83–0.85)
	ECOG PS 3	0.97 (0.96–0.98)	0.95 (0.92–0.98)	0.69 (0.63–0.75)	0.80 (0.79–0.81)
	ECOG PS 0–1	0.90 (0.89–0.91)	0.79 (0.76–0.82)	0.89 (0.87–0.91)	0.84 (0.83–0.85)
	ECOG PS 2–4	0.92 (0.91–0.93)	0.94 (0.92–0.96)	0.64 (0.60–0.68)	0.76 (0.75–0.77)

^a Accuracy including Unknown ECOG; ^b Training set included patients with ECOG PS 4 scores, however due to low performance, those ECOG PS 4 patients were grouped into Unknown ECOG Category for the results.

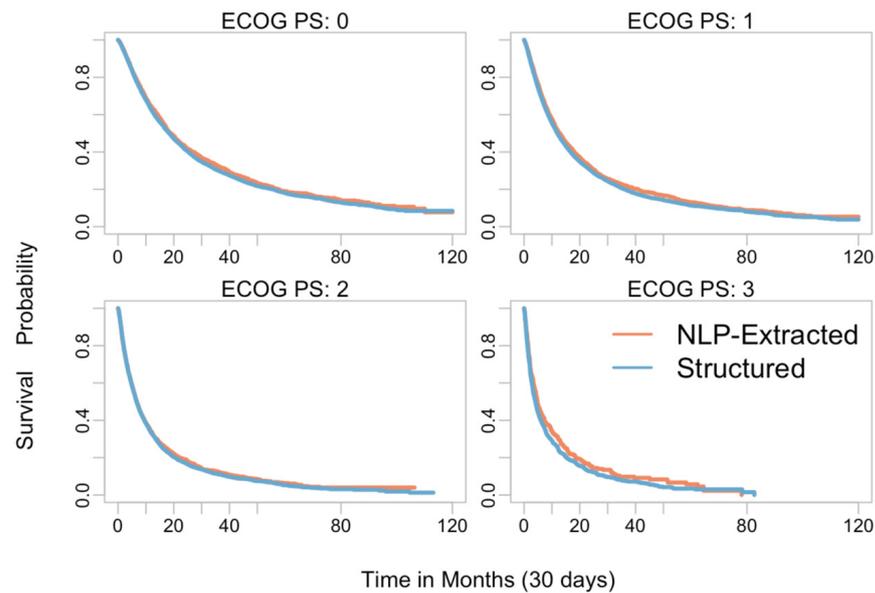


Figure 2. rwOS in patients with aNSCLC stratified according to their ECOG PS score for the subcohort with ECOG PS scores available as structured data, and for the subcohort with ECOG PS scores extracted via algorithm.

Likewise, an analysis looking at patients across all diseases in the testing set (stratified by ECOG PS score) showed similar results. Patients with worse ECOG PS lived less long than patients with better ECOG PS for both manually abstracted and extracted ECOG PS scores (26.7 to 5.8 months for patients with ECOG from abstraction and 26.4 to 4.3 months for patients with ECOG from NLP extraction) (Table S5).

5. Discussion

We developed an NLP algorithm that extracts ECOG PS scores from unstructured EHR documents with high performance compared to manually abstracted data. In contrast to prior research, which largely trained and applied ECOG extraction efforts to smaller single-site cohorts, this study applies a rule-based information extraction approach on a large nationwide EHR-derived de-identified RWD source to extract ECOG PS across a multitude of cancer types [44]. We showed that this strategy can achieve high performance, extracting ECOG PS at the time of treatment initiation, with a transparent and explainable algorithm deployed onto a large database that originated from real-world EHR data sources. The underlying study network from which the training and testing cohorts were selected includes approximately 280 cancer clinics (~800 sites of care) distributed nationwide (US-based), including academic and community practices, with a wide range of sizes [13,49]. The Flatiron Health database has been compared to The Surveillance, Epidemiology, and End Results Program (SEER) program and the National Program of Cancer Registries (NPCR)—both authoritative sources for population cancer surveillance and research in the US—and found general similarities in demographic and geographic distribution. However, patients from the Flatiron Health database appeared to be diagnosed with later stages of disease and their age distribution differs from the other datasets [49]. Another unique strength of this study is the availability of a patient sub-cohort with manually abstracted ECOG PS scores as the internal reference for quality, together with access to the primary EHR data sources, both structured and unstructured. This infrastructure enabled the verification of discrepancies and the identification and investigation of potential errors in both directions (of both the NLP algorithm-extracted and the manually-abstracted data).

Another strength of this study includes our improvement upon data completeness for EHR-derived study cohorts in oncology. In our database of interest ($N = 480,825$ patient-lines), use of this algorithm increased overall ECOG PS completeness from 60.4% to 73.2%. Guidelines from health authorities have highlighted completeness as a key component of data integrity and

a common shortcoming of observational data [51,52], spurring investigation in this area. NLP has been proven useful for the extraction of information from clinical notes in oncology EHRs; for example, date or site of disease recurrence, presence of symptoms or adverse events, date of disease onset, etcetera [28,30–33,40,53,54]. NLP techniques can be deployed to derive variables from EHR data, increasing data completeness and quality, boosting analytic robustness, and therefore improving the utility of RWE [18]. For our work, we considered NLP to be the best suited to extract explicit assessments of a patient's ECOG PS recorded in the EHR, rather than relying on contextual inferences made from EHR documents. The fact that NLP is traceable and auditable to source documentation is another benefit of the approach.

Furthermore, a strength of this study was our RWE focus in algorithmic evaluation. First, we stratified performance by ECOG score to allow researchers to understand the implications of using different ECOG PS-based criteria in their studies. Additionally, we assessed how NLP-extracted ECOG PS scores correlate with overall survival. Ensuring that extracted ECOG PS scores align with clinical expectations (the observation that patients with worse ECOG PS scores do not live as long [55–59]) provides reassurance about the prognostic value of the data extracted by the algorithm. By evaluating both the prognostic value as well as reporting out metrics such as sensitivity and PPV and utilizing error analysis, researchers can be surer of the impact of using these extracted clinical details in generating RWE. A final strength of this study is the ability to extract multiple ECOG PS scores for a patient over time and not just at time of diagnosis. This enables a better understanding of a patient's longitudinal journey over time, as well as improved completeness of ECOG PS in later lines of therapy, which may be important for some research studies focused on care provided towards the end of life.

Given the relevance of ECOG PS as a variable in oncology research, and the challenges associated with its overall documentation in RWD sources, our work has several implications for investigators. ECOG PS is critical to define potential clinical trial eligibility for use-cases in which RWE is generated for contextualization of clinical trials results [60,61]. ECOG PS incompleteness may impair the generation of adequate study samples and introduce bias. Therefore, maximizing availability of this variable would facilitate these contextual studies and would minimize any potential selection bias associated with reliance only on patients with structured ECOG PS. ECOG PS can be an important confounder in oncology studies, the application of this algorithm can increase analytic robustness, interpretability, and relevance by enabling multivariate analyses, increasing their statistical power, and by mitigating the need for imputation approaches altogether. Similar NLP algorithms could be deployed in practice, beyond analytic scenarios, to automate clinical trial eligibility screening processes or to enhance risk assessment to aid in clinical decision making.

While understanding the generalizability to other tumor types warrants further research, we believe basing our characterization step for the variable and the deeper analyses on one solid tumor type aNSCLC is an important contribution to the field and, to our knowledge, is among the first of its kind. For the scope of the present study, aNSCLC provided a disease setting where the ECOG PS documentation is particularly useful (due to expected frequent treatment changes in a short disease course), and the investigation of its relevance has been thorough [36,62,63], providing an ample backdrop of scientific literature to ascertain the performance of our NLP-extracted variable. Furthermore, while completeness of ECOG PS is indeed improved using an NLP algorithm, many patients still do not have an explicitly documented ECOG PS score and ECOG PS 3–4 is infrequently recorded. Future research should focus on novel methods to infer functional status and improve our understanding of the health of patients missing ECOG PS scores in their records.

Limitations

Study limitations include the variability of our algorithm's performance across settings and across score strata. This may be due to the relatively small sub-cohort sizes for the

low-prevalence strata (usually the poorer ECOG PS scores), or the potentially different documentation practices according to ECOG PS score or other clinician/patient factors. It is important to reiterate here that we did not categorize ECOG PS4, but grouped it into the Unknown category due to rare prevalence. This resulted in low performance of the algorithm. Nonetheless, the use of ECOG PS scores in clinical studies is most often based on the cutoff 0–1 vs. >1; therefore, the performance of this algorithm for this binary categorization is valuable. Furthermore, the prognostic analysis also boosts the confidence on the meaningfulness of the scores extracted by the algorithm, since they perform as would be expected. Finally, the algorithm itself was developed for the specific EHR-derived data source used; as such, further research will be needed to develop algorithms for different data sources.

6. Conclusions

In conclusion, we developed an NLP algorithm that extracts ECOG PS scores from unstructured documents across an EHR network of diverse real-world clinical settings. The algorithm improves the completeness of a critical variable at the time of treatment initiation and performs well across 21 distinct cancer types. Furthermore, additional validation, including the comparison of rwOS of patients with extracted vs. structured ECOG PS scores within a cohort of patients with aNSCLC, demonstrated clinically expected results. The use of high-performing algorithms such as this can help to overcome key challenges in RWD, such as missingness, as well as make one of the key advantages of RWD more attainable: the capability to aggregate longitudinal routine clinical care information from large patient cohorts for high-quality clinical research, ultimately benefiting providers, regulatory stakeholders, and, most importantly, patients.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/app13106209/s1>, Figure S1: rwOS in patients present in the study databases across all eligible diseases, stratified according to their ECOG PS score; Table S1: Detailed cohort eligibility criteria; Table S2: Impact of the application of the NLP algorithm on the ECOG PS completeness in EHR-derived databases for 21 diseases; Table S3: rwOS (months) in patients with aNSCLC stratified according to their ECOG PS score for the subcohort with ECOG PS scores available as structured data, and for the subcohort with ECOG PS scores extracted via algorithm; Table S4: HR for patients with structured ECOG (reference group) in patients with aNSCLC. Table S5. Median real world OS (months) for patients present in the testing set across all eligible diseases.

Author Contributions: Concept and design: A.B.C., C.J., A.R., K.H., M.R., S.N., G.A., B.H. and R.M. Data collection, analysis and interpretation: A.B.C., C.J., A.R., K.H., M.R., S.N. and B.H. Manuscript writing and review: All. 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 member 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 study 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: Requests for data sharing by license/permission for the purpose of replicating results in this manuscript can be submitted to dataaccess@flatiron.com.

Acknowledgments: Editorial support was provided by Julia Saiz, Hannah Gilham and Jennifer Swanson of Flatiron Health.

Conflicts of Interest: A.B.C., C.J., A.R., K.H., M.R., S.N., G.A. and R.M. report employment in Flatiron Health Inc., which is an independent subsidiary of the Roche Group, and stock ownership in Roche. R.M. and A.B.C. report equity ownership in Flatiron Health. B.H. reports consulting fees for Flatiron Health, the National Kidney Foundation, and Value Analytics Labs.

References

1. Khozin, S.; Blumenthal, G.; Pazdur, R. Real-world Data for Clinical Evidence Generation in Oncology. *J. Natl. Cancer Inst.* **2017**, *109*, djx187. [[CrossRef](#)] [[PubMed](#)]
2. Berger, M.L.; Curtis, M.D.; Smith, G.; Harnett, J.; Abernethy, A.P. Opportunities and challenges in leveraging electronic health record data in oncology. *Future Oncol.* **2016**, *12*, 1261–1274. [[CrossRef](#)] [[PubMed](#)]
3. Callahan, A.; Shah, N.H.; Chen, J.H. Research and Reporting Considerations for Observational Studies Using Electronic Health Record Data. *Ann. Intern. Med.* **2020**, *172* (Suppl. S11), S79–S84. [[CrossRef](#)] [[PubMed](#)]
4. Rudin, R.S.; Friedberg, M.W.; Shekelle, P.; Shah, N.; Bates, D.W. Getting Value from Electronic Health Records: Research Needed to Improve Practice. *Ann. Intern. Med.* **2020**, *172* (Suppl. S11), S130–S136. [[CrossRef](#)]
5. Guinn, D.; Wilhelm, E.E.; Lieberman, G.; Khozin, S. Assessing function of electronic health records for real-world data generation. *BMJ Evid.-Based Med.* **2018**, *24*, 95–98. [[CrossRef](#)]
6. 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](#)]
7. Tayefi, M.; Ngo, P.; Chomutare, T.; Dalianis, H.; Salvi, E.; Budrionis, A.; Godtlielsen, F. Challenges and opportunities beyond structured data in analysis of electronic health records. *WIREs Comput. Stat.* **2021**, *13*, e1549. [[CrossRef](#)]
8. Beaulieu-Jones, B.K.; Lavage, D.R.; Snyder, J.W.; Moore, J.H.; Pendergrass, S.A.; Bauer, C.R. Characterizing and Managing Missing Structured Data in Electronic Health Records: Data Analysis. *JMIR Public Health Surveill.* **2018**, *6*, e11. [[CrossRef](#)]
9. Perkins, N.; Cole, S.R.; Harel, O.; Tchetgen, E.J.T.; Sun, B.; Mitchell, E.M.; Schisterman, E. Principled Approaches to Missing Data in Epidemiologic Studies. *Am. J. Epidemiol.* **2017**, *187*, 568–575. [[CrossRef](#)]
10. Haneuse, S.; Arterburn, D.; Daniels, M.J. Assessing Missing Data Assumptions in EHR-Based Studies: A Complex and Underappreciated Task. *JAMA Netw. Open* **2021**, *4*, e210184. [[CrossRef](#)]
11. Kruse, C.S.; Stein, A.; Thomas, H.; Kaur, H. The use of Electronic Health Records to Support Population Health: A Systematic Review of the Literature. *J. Med. Syst.* **2018**, *42*, 1–16. [[CrossRef](#)] [[PubMed](#)]
12. Zhao, Y.; Howard, R.; Amorrortu, R.P.; Stewart, S.C.; Wang, X.; Calip, G.S.; Rollison, D.E. Assessing the Contribution of Scanned Outside Documents to the Completeness of Real-World Data Abstraction. *JCO Clin. Cancer Inform.* **2023**, *7*, e2200118. [[CrossRef](#)]
13. 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.
14. Waskom, M.L.; Tan, K.; Wiberg, H.; Cohen, A.B.; Wittmershaus, B.; Shapiro, W. A hybrid approach to scalable real-world data curation by machine learning and human experts. *medRxiv* **2023**. [[CrossRef](#)]
15. Bhattad, P.B.; Jain, V. Artificial Intelligence in Modern Medicine—The Evolving Necessity of the Present and Role in Transforming the Future of Medical Care. *Cureus* **2020**, *12*, e8041. [[CrossRef](#)]
16. Esteva, A.; Robicquet, A.; Ramsundar, B.; Kuleshov, V.; Depristo, M.; Chou, K.; Cui, C.; Corrado, G.; Thrun, S.; Dean, J. A guide to deep learning in healthcare. *Nat. Med.* **2019**, *25*, 24–29. [[CrossRef](#)]
17. Rajkumar, A.; Dean, J.; Kohane, I. Machine Learning in Medicine. *N. Engl. J. Med.* **2019**, *380*, 1347–1358. [[CrossRef](#)]
18. Datta, S.; Bernstam, E.V.; Roberts, K. A frame semantic overview of NLP-based information extraction for cancer-related EHR notes. *J. Biomed. Inform.* **2019**, *100*, 103301. [[CrossRef](#)]
19. Bertsimas, D.; Wiberg, H. Machine Learning in Oncology: Methods, Applications, and Challenges. *JCO Clin. Cancer Inform.* **2020**, *4*, 885–894. [[CrossRef](#)]
20. Yim, W.-W.; Yetisgen, M.; Harris, W.P.; Kwan, S.W. Natural Language Processing in Oncology. *JAMA Oncol.* **2016**, *2*, 797–804. [[CrossRef](#)]
21. Rajman, M.; Besançon, R. Text Mining: Natural Language techniques and Text Mining applications. In *Data Mining and Reverse Engineering. IFIP—The International Federation for Information Processing*; Spaccapietra, S., Maryanski, F., Eds.; Springer: Boston, MA, USA, 1998; pp. 50–64. [[CrossRef](#)]
22. Bera, K.; Schalper, K.A.; Rimm, D.L.; Velcheti, V.; Madabhushi, A. Artificial intelligence in digital pathology—New tools for diagnosis and precision oncology. *Nat. Rev. Clin. Oncol.* **2019**, *16*, 703–715. [[CrossRef](#)] [[PubMed](#)]
23. Feeny, A.K.; Chung, M.K.; Madabhushi, A.; Attia, Z.L.; Cikes, M.; Firouznia, M.; Friedman, P.A.; Kalscheur, M.M.; Kapa, S.; Narayan, S.M.; et al. Artificial Intelligence and Machine Learning in Arrhythmias and Cardiac Electrophysiology. *Circ. Arrhythmia Electrophysiol.* **2020**, *13*, e007952. [[CrossRef](#)] [[PubMed](#)]
24. D'amore, B.; Smolinski-Zhao, S.; Daye, D.; Uppot, R.N. Role of Machine Learning and Artificial Intelligence in Interventional Oncology. *Curr. Oncol. Rep.* **2021**, *23*, 1–8. [[CrossRef](#)]
25. Jayatilake, S.M.D.A.C.; Ganegoda, G.U. Involvement of Machine Learning Tools in Healthcare Decision Making. *J. Healthc. Eng.* **2021**, *2021*, 6679512. [[CrossRef](#)] [[PubMed](#)]
26. Jiang, M.; Ma, Y.; Guo, S.; Jin, L.; Lv, L.; Han, L.; An, N. Using Machine Learning Technologies in Pressure Injury Management: Systematic Review. *JMIR Public Health Surveill.* **2021**, *9*, e25704. [[CrossRef](#)] [[PubMed](#)]
27. Peterson, D.J.; Ostberg, N.P.; Blayney, D.W.; Brooks, J.D.; Hernandez-Boussard, T. Machine Learning Applied to Electronic Health Records: Identification of Chemotherapy Patients at High Risk for Preventable Emergency Department Visits and Hospital Admissions. *JCO Clin. Cancer Inform.* **2021**, *5*, 1106–1126. [[CrossRef](#)]

28. Banerjee, I.; Bozkurt, S.; Caswell-Jin, J.; Kurian, A.W.; Rubin, D.L. Natural Language Processing Approaches to Detect the Timeline of Metastatic Recurrence of Breast Cancer. *JCO Clin. Cancer Inform.* **2019**, *3*, 1–12. [[CrossRef](#)] [[PubMed](#)]
29. Karimi, Y.H.; Blayney, D.W.; Kurian, A.W.; Shen, J.; Yamashita, R.; Rubin, D.; Banerjee, I. Development and Use of Natural Language Processing for Identification of Distant Cancer Recurrence and Sites of Distant Recurrence Using Unstructured Electronic Health Record Data. *JCO Clin. Cancer Inform.* **2021**, *5*, 469–478. [[CrossRef](#)]
30. Kehl, K.L.; Xu, W.; Lepisto, E.; Elmarakeby, H.; Hassett, M.J.; Van Allen, E.M.; Johnson, B.E.; Schrag, D. Natural Language Processing to Ascertain Cancer Outcomes from Medical Oncologist Notes. *JCO Clin. Cancer Inform.* **2020**, *4*, 680–690. [[CrossRef](#)]
31. Fu, S.; Chen, D.; He, H.; Liu, S.; Moon, S.; Peterson, K.J.; Shen, F.; Wang, L.; Wang, Y.; Wen, A.; et al. Clinical concept extraction: A methodology review. *J. Biomed. Inform.* **2020**, *109*, 103526. [[CrossRef](#)]
32. Savova, G.K.; Danciu, I.; Alamudun, F.; Miller, T.; Lin, C.; Bitterman, D.S.; Tourassi, G.; Warner, J.L. Use of Natural Language Processing to Extract Clinical Cancer Phenotypes from Electronic Medical Records. *Cancer Res* **2019**, *79*, 5463–5470. [[CrossRef](#)] [[PubMed](#)]
33. Deshmukh, P.R.; Phalnikar, R. Information extraction for prognostic stage prediction from breast cancer medical records using NLP and ML. *Med. Biol. Eng. Comput.* **2021**, *59*, 1751–1772. [[CrossRef](#)] [[PubMed](#)]
34. Oken, M.M.; Creech, R.H.; Tormey, D.C.; Horton, J.; Davis, T.E.; McFadden, E.T.; Carbone, P.P. Toxicity and response criteria of the Eastern Cooperative Oncology Group. *Am. J. Clin. Oncol.* **1982**, *5*, 649–656. [[CrossRef](#)] [[PubMed](#)]
35. Albain, K.S.; Crowley, J.J.; Leblanc, M.; Livingston, R.B. Survival determinants in extensive-stage non-small-cell lung cancer: The Southwest Oncology Group experience. *J. Clin. Oncol.* **1991**, *9*, 1618–1626. [[CrossRef](#)] [[PubMed](#)]
36. Jang, R.W.; Caraiscos, V.B.; Swami, N.; Banerjee, S.; Mak, E.; Kaya, E.; Rodin, G.; Bryson, J.; Ridley, J.Z.; Le, L.W.; et al. Simple Prognostic Model for Patients with Advanced Cancer Based on Performance Status. *J. Oncol. Pr.* **2014**, *10*, e335–e341. [[CrossRef](#)]
37. Köhne, C.-H.; Cunningham, D.; Di Costanzo, F.; Glimelius, B.; Blijham, G.; Aranda, E.; Scheithauer, W.; Rougier, P.; Palmer, M.; Wils, J.; et al. Clinical determinants of survival in patients with 5-fluorouracil-based treatment for metastatic colorectal cancer: Results of a multivariate analysis of 3825 patients. *Ann. Oncol.* **2002**, *13*, 308–317. [[CrossRef](#)]
38. Sargent, D.J.; Köhne, C.H.; Sanoff, H.K.; Bot, B.M.; Seymour, M.T.; de Gramont, A.; Porschen, R.; Saltz, L.B.; Rougier, P.; Tournigand, C.; et al. Pooled Safety and Efficacy Analysis Examining the Effect of Performance Status on Outcomes in Nine First-Line Treatment Trials Using Individual Data from Patients with Metastatic Colorectal Cancer. *J. Clin. Oncol.* **2009**, *27*, 1948–1955. [[CrossRef](#)]
39. Schiller, J.H.; Harrington, D.; Belani, C.P.; Langer, C.; Sandler, A.; Krook, J.; Zhu, J.; Johnson, D.H. Comparison of Four Chemotherapy Regimens for Advanced Non-Small-Cell Lung Cancer. *N. Engl. J. Med.* **2002**, *346*, 92–98. [[CrossRef](#)]
40. Sengeløv, L.; Kamby, C.; Geertsens, P.; Andersen, L.J.; von der Maase, H. Predictive factors of response to cisplatin-based chemotherapy and the relation of response to survival in patients with metastatic urothelial cancer. *Cancer Chemother. Pharmacol.* **2000**, *46*, 357–364. [[CrossRef](#)]
41. Blagden, S.P.; Charman, S.C.; Sharples, L.D.; Magee, L.R.A.; Gilligan, D. Performance status score: Do patients and their oncologists agree? *Br. J. Cancer* **2003**, *89*, 1022–1027. [[CrossRef](#)]
42. Roila, F.; Lupattelli, M.; Sassi, M.; Basurto, C.; Bracarda, S.; Picciafuoco, M.; Boschetti, E.; Milella, G.; Ballatori, E.; Tonato, M.; et al. Intra and interobserver variability in cancer patients' performance status assessed according to Karnofsky and ECOG scales. *Ann. Oncol.* **1991**, *2*, 437–439. [[CrossRef](#)] [[PubMed](#)]
43. Sorensen, J.B.; Klee, M.R.; Palshof, T.; Hansen, H.H. Performance status assessment in cancer patients. An inter-observer variability study. *Br. J. Cancer* **1993**, *67*, 773–775. [[CrossRef](#)] [[PubMed](#)]
44. Wang, L.; Fu, S.; Wen, A.; Ruan, X.; He, H.; Liu, S.; Moon, S.; Mai, M.; Riaz, I.B.; Wang, N.; et al. Assessment of Electronic Health Record for Cancer Research and Patient Care Through a Scoping Review of Cancer Natural Language Processing. *JCO Clin. Cancer Inform.* **2022**, *6*, e2200006. [[CrossRef](#)] [[PubMed](#)]
45. Hom, J.; Nikowitz, J.; Ottesen, R.; Niland, J.C. Facilitating clinical research through automation: Combining optical character recognition with natural language processing. *Clin. Trials* **2022**, *19*, 504–511. [[CrossRef](#)]
46. Agarannik, N.; Lindvall, C.; El-Jawahri, A.; He, W.; Iezzoni, L. Use of Natural Language Processing to Assess Frequency of Functional Status Documentation for Patients Newly Diagnosed with Colorectal Cancer. *JAMA Oncol.* **2020**, *6*, 1628–1630. [[CrossRef](#)]
47. Gauthier, M.-P.; Law, J.H.; Le, L.W.; Li, J.J.; Zahir, S.; Nirmalakumar, S.; Sung, M.; Pettengell, C.; Aviv, S.; Chu, R.; et al. Automating Access to Real-World Evidence. *JTO Clin. Res. Rep.* **2022**, *3*, 100340. [[CrossRef](#)]
48. Herath, D.H.; Wilson-Ing, D.; Ramos, E.; Morstyn, G. Assessing the natural language processing capabilities of IBM Watson for oncology using real Australian lung cancer cases. *J. Clin. Oncol.* **2016**, *34*, e18229. [[CrossRef](#)]
49. Ma, X.; Long, L.; Moon, S.; Adamson, B.J.; Baxi, S.S. Comparison of Population Characteristics in Real-World Clinical Oncology Databases in the US: Flatiron Health, SEER, and NPCR. *medRxiv* **2020**. [[CrossRef](#)]
50. Haimson, J.D.; Baxi, S.; Meropol, N.; Ambwani, G.; Backenroth, D.; Murali, M.; Rosic, A.; Chengsheng, J. Prognostic Score Based on Health Information. U.S. Patent 11651252, 16 May 2023.
51. Center for Drug Evaluation and Research Center for Biologics Evaluation and Research Oncology Center of Excellence. 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. US Food & Drug Administration Web Site. September 2021. Available online: <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-electronic-health-records-and-medical-claims-data-support-regulatory> (accessed on 27 February 2023).

52. Kent, S.; Burn, E.; Dawoud, D.; Jonsson, P.; Østby, J.T.; Hughes, N.; Rijnbeek, P.; Bouvy, J.C. Common Problems, Common Data Model Solutions: Evidence Generation for Health Technology Assessment. *Pharmacoeconomics* **2020**, *39*, 275–285. [[CrossRef](#)]
53. Gupta, S.; Belouali, A.; Shah, N.J.; Atkins, M.B.; Madhavan, S. Automated Identification of Patients with Immune-Related Adverse Events from Clinical Notes Using Word Embedding and Machine Learning. *JCO Clin. Cancer Inform.* **2021**, *5*, 541–549. [[CrossRef](#)]
54. Koleck, T.A.; Dreisbach, C.; Bourne, P.E.; Bakken, S. Natural language processing of symptoms documented in free-text narratives of electronic health records: A systematic review. *J. Am. Med. Inform. Assoc.* **2019**, *26*, 364–379. [[CrossRef](#)] [[PubMed](#)]
55. Dall’olio, F.G.; Maggio, I.; Massucci, M.; Mollica, V.; Fragonomeno, B.; Ardizzoni, A. ECOG performance status ≥ 2 as a prognostic factor in patients with advanced non small cell lung cancer treated with immune checkpoint inhibitors—A systematic review and meta-analysis of real world data. *Lung Cancer* **2020**, *145*, 95–104. [[CrossRef](#)] [[PubMed](#)]
56. Kawaguchi, T.; Takada, M.; Kubo, A.; Matsumura, A.; Fukai, S.; Tamura, A.; Saito, R.; Maruyama, Y.; Kawahara, M.; Ou, S.-H.I. Performance Status and Smoking Status Are Independent Favorable Prognostic Factors for Survival in Non-small Cell Lung Cancer: A Comprehensive Analysis of 26,957 Patients with NSCLC. *J. Thorac. Oncol.* **2010**, *5*, 620–630. [[CrossRef](#)]
57. Kawsar, H.; Gaudel, P.; Suleiman, N.; Al-Jumayli, M.; Huang, C.; Neupane, P. 221 Poor performance status negatively affects survival benefit of immunotherapy in non-small cell lung cancer. *J. Immunother. Cancer* **2020**, *8*, A131–A132. [[CrossRef](#)]
58. Sehgal, K.; Gill, R.R.; Widick, P.; Bindal, P.; McDonald, D.C.; Shea, M.; Rangachari, D.; Costa, D.B. Association of Performance Status with Survival in Patients with Advanced Non-Small Cell Lung Cancer Treated With Pembrolizumab Monotherapy. *JAMA Netw. Open* **2021**, *4*, e2037120. [[CrossRef](#)]
59. Catalano, M.; Aprile, G.; Conca, R.; Petrioli, R.; Ramello, M.; Roviello, G. The impact of age, performance status and comorbidities on nab-paclitaxel plus gemcitabine effectiveness in patients with metastatic pancreatic cancer. *Sci. Rep.* **2022**, *12* (Suppl. S3), 1–7. [[CrossRef](#)]
60. Petito, L.; García-Albéniz, X.; Logan, R.W.; Howlader, N.; Mariotto, A.B.; Dahabreh, I.J.; Hernán, M.A. Estimates of Overall Survival in Patients with Cancer Receiving Different Treatment Regimens: Emulating Hypothetical Target Trials in the Surveillance, Epidemiology, and End Results (SEER)-Medicare Linked Database. *JAMA Netw. Open* **2020**, *3*, e200452. [[CrossRef](#)]
61. Tan, K.; Bryan, J.; Segal, B.; Bellomo, L.; Nussbaum, N.; Tucker, M.; Torres, A.Z.; Bennette, C.; Capra, W.; Curtis, M.; et al. Emulating Control Arms for Cancer Clinical Trials Using External Cohorts Created from Electronic Health Record-Derived Real-World Data. *Clin. Pharmacol. Ther.* **2021**, *111*, 168–178. [[CrossRef](#)]
62. Lilenbaum, R.C.; Cashy, J.; Hensing, T.A.; Young, S.; Cella, D. Prevalence of Poor Performance Status in Lung Cancer Patients: Implications for Research. *J. Thorac. Oncol.* **2008**, *3*, 125–129. [[CrossRef](#)]
63. Boukovinas, I.; Kosmidis, P. Treatment of non-small cell lung cancer patients with performance status2 (PS2). *Lung Cancer* **2009**, *63*, 10–15. [[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.