original reports abstract

Efficient and Accurate Extracting of Unstructured EHRs on Cancer Therapy Responses for the Development of RECIST Natural Language Processing Tools: Part I, the Corpus

Yalun Li, MD[1,2]; Yung-Hung Luo, MD[3]; Jason A. Wampfler, BS[4]; Samuel M. Rubinstein, MD[5]; Firat Tiryaki, BSc[6]; Kumar Ashok V, BA[7]; Jeremy L. Warner, MD, MS[5,8]; Hua Xu, PhD[6]; and Ping Yang, MD[1]

PURPOSE Electronic health records (EHRs) are created primarily for nonresearch purposes; thus, the amounts of data are enormous, and the data are crude, heterogeneous, incomplete, and largely unstructured, presenting challenges to effective analyses for timely, reliable results. Particularly, research dealing with clinical notes relevant to patient care and outcome is seldom conducted, due to the complexity of data extraction and accurate annotation in the past. RECIST is a set of widely accepted research criteria to evaluate tumor response in patients undergoing antineoplastic therapy. The aim for this study was to identify textual sources for RECIST information in EHRs and to develop a corpus of pharmacotherapy and response entities for development of natural language processing tools.

METHODS We focused on pharmacotherapies and patient responses, using 55,120 medical notes (n = 72 types) in Mayo Clinic's EHRs from 622 randomly selected patients who signed authorization for research. Using the Multidocument Annotation Environment tool, we applied and evaluated predefined keywords, and time interval and note-type filters for identifying RECIST information and established a gold standard data set for patient outcome research.

RESULTS Key words reduced clinical notes to 37,406, and using four note types within 12 months postdiagnosis further reduced the number of notes to 5,005 that were manually annotated, which covered 97.9% of all cases (n = 609 of 622). The resulting data set of 609 cases (n = 503 for training and n = 106 for validation purpose), contains 736 fully annotated, deidentified clinical notes, with pharmacotherapies and four response end points: complete response, partial response, stable disease, and progressive disease. This resource is readily expandable to specific drugs, regimens, and most solid tumors.

CONCLUSION We have established a gold standard data set to accommodate development of biomedical informatics tools in accelerating research into antineoplastic therapeutic response.

JCO Clin Cancer Inform 4:383-391. © 2020 by American Society of Clinical Oncology

INTRODUCTION

An electronic health record (EHR) is a digital form of a patient's medical history, making real-time, patient-centered information available instantly and securely to authorized users. Although primarily designed for billing and to document medical and treatment histories of patients, EHR data can be used for other purposes. Recently, EHRs have been successfully implemented in the majority of US health care systems in various platforms, enabling a surge in secondary uses of EHRs, especially for research. However, because EHRs are created primarily for nonresearch purposes, derived data sets are enormous, crude, heterogeneous, incomplete, and largely unstructured, presenting challenges to effective analyses for timely new and reliable findings. Particularly, research dealing with unstructured clinical notes relevant to patient care and outcome, such as response to therapy, had been ineffectively conducted before the era of artificial intelligence (AI) techniques, due to the complexity of data extraction and accurate annotation. These challenges are being surmounted with the application of AI techniques (eg, clinical natural language processing [NLP] tools and machine learning)[1,2]; as a consequence, EHRs are gradually being used to facilitate and accelerate research relevant to patient care.[3]

RECIST[4] is a set of widely accepted rules to define tumor response for patients undergoing antineoplastic therapy. The majority of clinical trials evaluating cancer treatments for objective response in solid tumors use RECIST, which was developed and published in 2000 (version 1.0) and subsequently updated in 2009 (version 1.1), defining four levels of objective tumor

ASSOCIATED CONTENT

Data Supplement

Author affiliations and support information (if applicable) appear at the end of this article.

Accepted on March 19, 2020 and published at ascopubs.org/journal/cci on May 4, 2020: DOI https://doi.org/10.1200/CCI.19.00147

CONTEXT

Key Objective

To identify textual sources for the widely used RECIST information in electronic health records (EHRs) and develop a fully annotated gold standard corpus of anti-lung cancer pharmacotherapy and response entities.

Knowledge Generated

A data corpus was developed with pharmacotherapies and the four RECIST response levels of complete response, partial response, stable disease and progressive disease, starting with a total of 55,120 medical notes. A gold standard corpus was built containing 609 randomly selected lung cancer cases (n = 503 for training and n = 106 for validation) and 736 fully annotated clinical notes (n = 617 for training and n = 119 for validation).

Relevance

This resource is readily useable to train natural language processing tools for predefined clinical and research applications. It can be expanded to extract information on specific drugs, treatment regimens, and adverse effects; and the algorithm and tools can be adapted to most solid tumors for which therapy responses are evaluated by RECIST.

response for antineoplastic therapies: complete response (CR), partial response (PR), stable disease (SD), and progressive disease (PD). The RECIST definition of pharmacotherapy response has been widely used for cancer clinical and epidemiologic studies.[5,6] Despite wide use for clinical trial end points, documentation of formal RECIST evaluation and results are less common in nonclinical trial, real-world settings.[7] Moreover, manual extraction of RECIST information from EHRs is time consuming; thus, informatics tools that can automatically extract RECIST are highly desirable.

Our ultimate goal is to develop NLP tools that can automatically extract an approximation of cancer treatment and outcome information from clinical notes, because much of such information is embedded in narrative documents in EHRs. Nevertheless, two challenges exist for developing RECIST NLP tools: (1) it is not clear how RECIST information is stored in EHRs and (2) there is a lack of standard annotated corpori for RECIST information extraction. To address these challenges, we conducted this study to identify textual sources for RECIST information in EHRs and to develop a corpus of pharmacotherapy and response entities for development of NLP tools.

Although our current work focuses on predefined treatments as a single group, "pharmacotherapy," the gold standard data set with built-in training and validation cases established is readily expandable to specific drugs and regimens; moreover, the RECIST tools to be developed on the basis of the data set will be applicable to solid tumors.

METHODS

Assembly of a Gold Standard Data Set

The patient base for the data set is an existing, 20-year, lung cancer cohort established in the Mayo Clinic Epidemiology and Genetics of Lung Cancer database.[8-10]

Inclusion. All patients were aged ≥18 years and were selected from a previously described, large cohort of patients with primary lung cancer.[8-10] Eligible patients had lung cancer diagnosis confirmed by histopathology or cytology and were treated at Mayo Clinic. Patients were treated with at least one cycle of one or more antineoplastic drugs and had a measured drug response. All patients consented to participate in the study or had given authorization to allow their medical record information reviewed for research. Patients without antineoplastic drug exposure and with incomplete EHRs were not eligible for the current study, as shown in Figure 1. Noted is the time window between 1997 and 2000, when Mayo Clinic transitioned from paper medical records to EHRs, with an anticipated 10% incomplete conversion. We identified all patients with primary lung cancer who received antineoplastic drug therapy at Mayo Clinic with a known response. A random selection of 660 patients from this cohort was assigned to the training set (n = 542) or the validation set (n = 118); our aim was to achieve 600 total cases in the data set. A final total of 622 patients met the inclusion criteria for the study (Fig 1).

Identifying Sources of RECIST Information

Key terms annotation procedure. The categorization of systemic therapy by intent and possible responses are described in Table 1. For palliative, adjuvant, and induction pharmacotherapy, we evaluated patients for CR, PR, SD, and PD. Although consolidation and maintenance pharmacotherapy deepens responses in some studies, for the purposes of this analysis, we only evaluated for disease stability or progression during this phase of care.

Note types within Mayo Clinic EHRs. The Mayo Clinic EHR system contains at least 31 types of clinical documentation, which are described in the Data Supplement. The purpose of this study was to extract response data from unstructured clinical documentation; thus, we abstracted only the clinical documents relevant to this analysis. These are labeled in the Mayo EHR as Subsequent Visit (SV), Consult (CON),

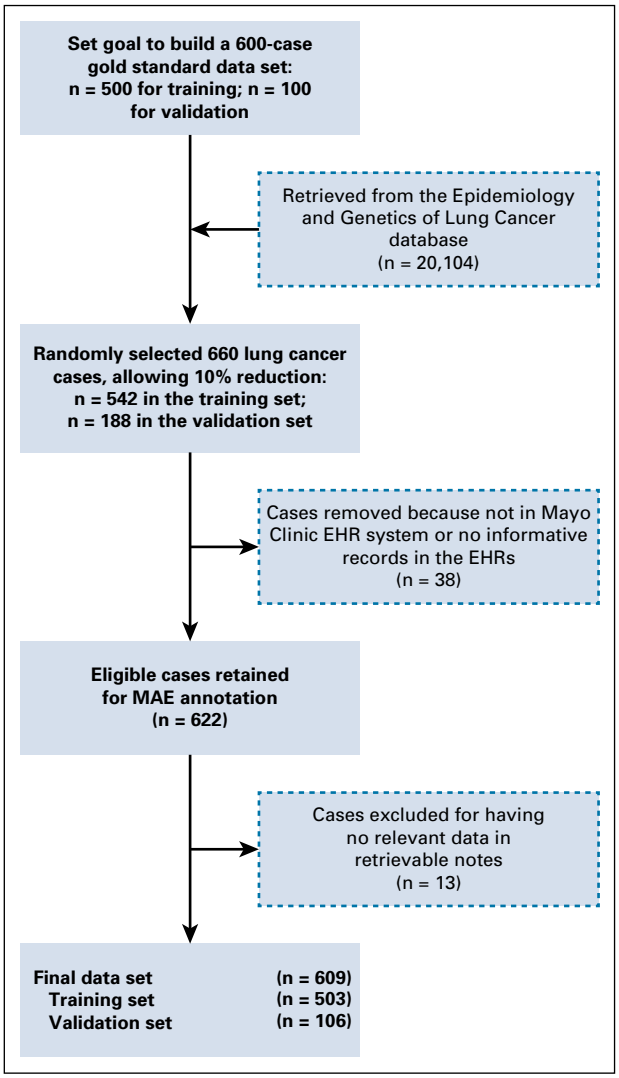

FIG 1. We randomly selected 660 patients, aiming to achieve 600 in the gold standard data set. A final total of 622 patients met the inclusion criteria for the study. The resulting data set comprised 609 cases, 503 for training and 106 for validation purposes. EHR, electronic health record; MAE, Multidocument Annotation Environment.

Miscellaneous (MIS), and Summary (SUM). There are many other conditions that are described within the EHRs that are common in patients with lung cancer and that may have relevance in the selection of therapy. These conditions were not currently annotated but can be extracted feasibly from the same documents used for the current work focusing on response assessment. We divided the conditions into three categories: comorbid diseases (comorbidity), drug-induced adverse conditions (drug-induced), or cancer-related symptoms not contributing to evaluation of RECIST responses to antineoplastic drugs (eg, symptomatic hypercalcemia).

Annotating RECIST Entities for Corpus Development

Annotation guidelines. Using an iterative approach, we developed guidelines for annotating RECIST information. Definitions of entities and attributes are described in the following paragraphs.

Definition of RECIST response. Taking as reference the smallest sum of diameters of target lesions, there are four response levels: (1) CR: Disappearance of all target lesions; any pathologic lymph nodes (whether target or nontarget) must have reduction in the short axis to < 10 mm. (2) PR: At least a 30% decrease in the sum of diameters of target lesions, taking as reference the baseline sum diameters. (3) PD: At least a 20% increase in the sum of diameters of target lesions (including baseline sum) and an absolute increase of at least 5 mm. The appearance of one or more new lesions is also considered progression. (4) SD: Neither sufficient shrinkage to qualify for PR nor sufficient increase to qualify for PD. Annotation of RECIST response was based on information found in the clinical notes summarizing the imaging findings, without the raw imaging read by coders themselves.

Definition of pharmacotherapy. An overview of lung cancer treatments and detailed pharmacotherapy strategies is provided in the Data Supplement. Our current work focuses on predefined treatments as a single group, pharmacologic drugs (Fig 2), and detailed drug information is provided in the Data Supplement. We documented details of alternative terms for pharmacotherapy in EHRs (ie, all encountered variations for therapy and responses) to account for the four RECIST categories. In addition, we also defined three attributes for CR, PR, PD, and SD:

1. Certainty: evidence of pharmacotherapy and evidence of each level of response. Negated: the problem does not exist in the EHRs. Possible: patient may have

TABLE 1. Systemic Therapies of Lung Cancer and Response Categories

		Response			
Systemic Therapy	**Disease Status**	**CR**	**PR**	**SD**	**PD**
Palliative pharmacotherapy	Advanced NSCLC	√	√	√	√
Adjuvant/induction pharmacotherapy	Completely resected NSCLC	√	√	√	√
Combination pharmacotherapy	SCLC	√	√	√	√
Maintenance/consolidative pharmacotherapy	Previously received drug therapy			√	√

Abbreviations: CR, complete response; NSCLC, non–small-cell lung cancer; PD, progressive disease; PR, partial response; SCLC, small-cell lung cancer; SD, stable disease.

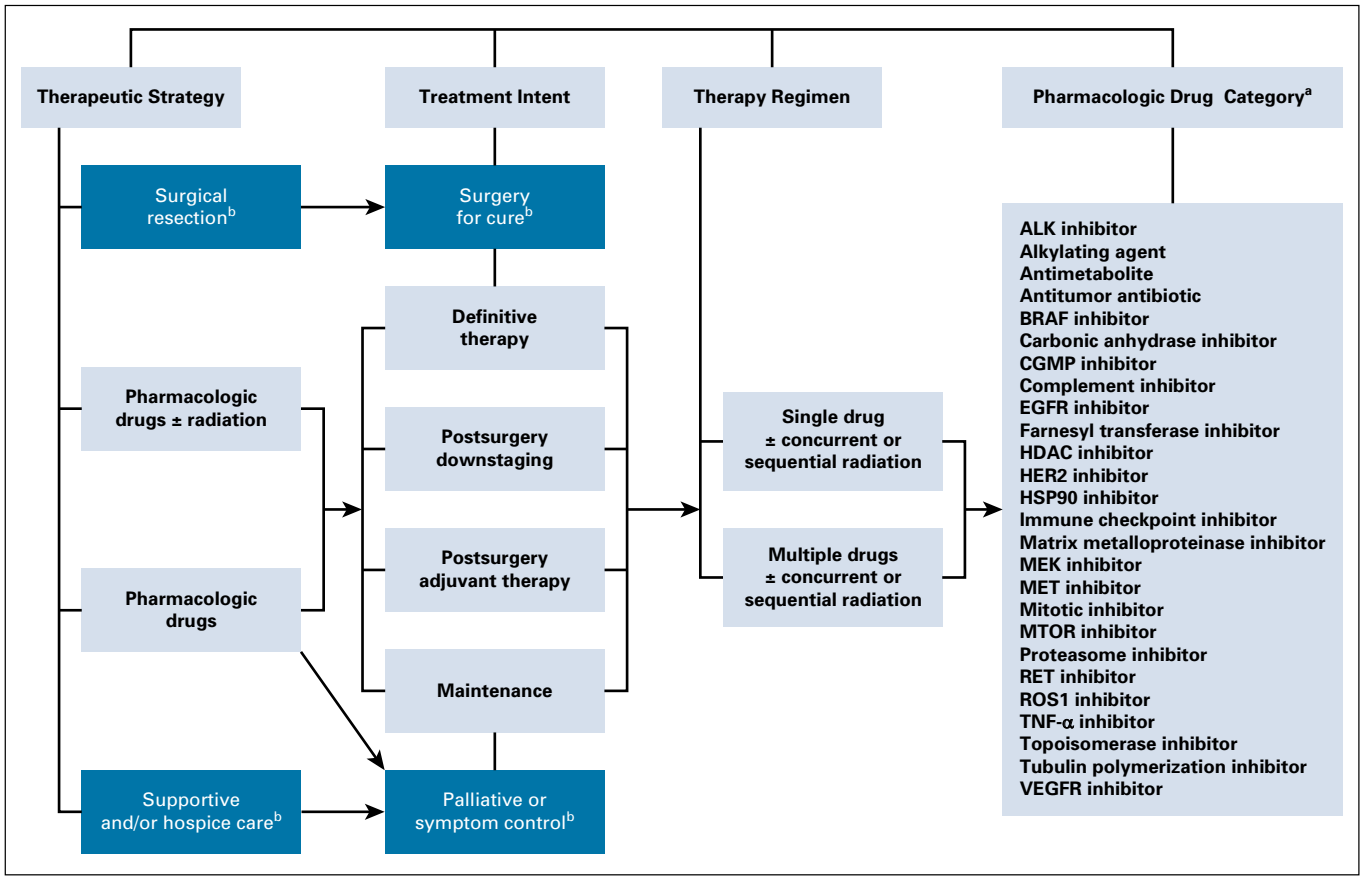

FIG 2. There are four major dimensions of antineoplastic treatment of primary and/or recurrent tumors: therapeutic strategy, treatment intents, therapy regimen, and pharmacologic drug category. ([a]) Alphabetically for coding search convenience. ([b]) Not in the scope of current study.

a problem, but there is uncertainty expressed in the note. Possible takes precedence over negated, so terms like "probably not" or "unlikely" categorize problems as being possible, just as "probably" and "likely" do. Hypothetical: medical problems the note asserts the patient may develop or discuss with. Positive: the chemotherapy or response was confirmed.

2. Status: the tense in which treatment or response occurs: historical or present
3. Exclusion: Key terms used outside of the study scope. Exclude the responses that are not for lung cancer or lung site or pharmacotherapy.

Annotation procedure. **Deidentification.** We used an in-house pipeline to detect protected health information (PHI) and manual annotation of PHI in parallel by two independent teams; subsequently, the PHI information identified through the pipeline and the manual way were combined and removed.

Annotation tool. Multidocument Annotation Environment (MAE), an annotation tool for natural language text annotation, allows users to define their own annotation tasks flexibly, annotate partial words, use nonconsuming tags, and easily create links between extent entities. It outputs annotations in stand-off XML. Although it does not enforce strict rules for annotation schemas, it is easy to set up and start running.[11] Although MAE does not represent a new frontier in annotation software, its ease of use, portability, and clean visualization make it useful and efficient for annotation projects.

Training and annotation procedures. Three coders were trained; coders 1 and 2 performed the entire annotation and coder 3 was an adjudicator when disagreement occurred between coders 1 and 2 during a five-step training process using five note sets: set 1: 10 notes (notes 1-10), learning together with a trained coder; set 2: 15 notes (notes 11-25), learning independently, dissecting, and checking; set 3: 11 notes (notes 26-36), coding independently, debating, and converging; set 4: 17 notes (notes 37-53), coding independently, comparing, and refining; and set 5: 28 notes (notes 54-71), coding independently and reaching consensus. Detailed training results are provided in the Data Supplement.

Annotated informative clinical notes. For most cases, several selected clinical notes contained duplicated information and stated the same period of pharmacotherapy or a measured drug response, largely were not directly

relevant to antineoplastic therapy, and instead, related to patients' comorbid conditions and treatment adverse effects or toxicities, as listed in the Data Supplement. To concentrate informative data during annotation, we searched, reviewed, and selected at least one clinical note for each case covering all key terms.

TABLE 2. Clinical Characteristics of Gold Standard Data-Set Cases

Clinical Characteristic	Training Set (n = 503)	Validation Set (n = 106)	Total (N = 609)	*P*
Age at diagnosis, years				.303
Mean (SD)	64.2 (10.4)	65.1 (10.5)	64.3 (10.4)	
Range	32.2-89.3	31.7-87.4	31.7-89.3	
Sex				.248
Female	273 (54.3)	51 (48.1)	324 (53.2)	
Male	230 (45.7)	55 (51.9)	285 (46.8)	
Smoker status				.958
Never	77 (15.3)	17 (16.0)	94 (15.4)	
Former	221 (43.9)	45 (42.5)	266 (43.7)	
Current/ever	205 (40.8)	44 (41.5)	249 (40.9)	
Lung cancer type				.823
NSCLC	332 (66.0)	72 (67.9)	404 (66.3)	
SCLC	160 (31.8)	31 (29.2)	191 (31.4)	
Other lung malignancy	11 (2.2)	3 (2.8)	14 (2.3)	
TNM stage: NSCLC[a]				.253
I-IIIA[b]	144 (42.1)	37 (49.3)	181 (43.4)	
IIIB-IV	198 (57.9)	38 (50.7)	236 (56.6)	
TNM stage: SCLC				.208
Limited	83 (52.2)	20 (64.5)	103 (54.2)	
Extensive	76 (47.8)	11 (35.5)	87 (45.8)	
Condensed grade				.422
Well differentiation	27 (5.4)	4 (3.8)	31 (5.1)	
Moderate differentiation	132 (26.2)	35 (33.0)	167 (27.4)	
Poor or no differentiation	247 (49.1)	45 (42.5)	292 (47.9)	
Nongradable	97 (19.3)	22 (20.8)	119 (19.5)	
Therapy				.658
Surgery and adjuvant chemotherapy	49 (9.7)	12 (11.3)	61 (10.0)	
Surgery and adjuvant chemoradiotherapy	75 (14.9)	11 (10.4)	86 (14.1)	
Chemotherapy only	158 (31.4)	34 (32.1)	192 (31.5)	
Chemoradiotherapy	221 (43.9)	49 (46.2)	270 (44.3)	
Response				.688
Complete response	59 (11.7)	16 (15.1)	75 (12.3)	
Partial response	193 (38.4)	36 (34.0)	229 (37.6)	
Stable	76 (15.1)	18 (17.0)	94 (15.4)	
Progression	175 (34.8)	36 (34.0)	211 (34.6)	
Follow-up				.759
Mean (SD)	35.5 (43.9)	32.4 (39.7)	35.0 (43.2)	
Median	18.1	18.0	18.1	
Range	1.2-242.4	2.2-226.8	1.2-242.4	

NOTE. Data reported as No. (%) unless otherwise indicated.

Abbreviations: NSCLC, non—small-cell lung cancer; SCLC, small-cell lung cancer.

[a]Three patients were diagnosed with lymphoma that could not be defined in TNM stage.

[b]Patients with early-stage (stage I/II) disease received adjuvant chemotherapy pre- or postsurgery.

RESULTS

We randomly selected 622 eligible patients with lung cancer in our database and extracted 55,120 total medical notes (n = 72 types) in our institutional EHR system. After applying the key words "complete response," "partial response," "stable," and "progressive disease," we reduced clinical notes to 37,406 in 31 types (Data Supplement). Then, using four note types (SV, CON, MIS, SUM) within 12 months after diagnosis further reduced the number of clinical notes to 5,005—9% of the original total (n = 5,005 of 55,120). Most cases (90%) had more than five notes before annotation (Data Supplement). Manual annotation using MAE narrowed 1% of cases to more than four notes; the resulting 736 notes included all key terms and covered 97.9% of all cases (n = 609 of 622); the remaining notes had repeated or irrelevant information to the current work. Detailed annotations of 1.3% of the total clinical notes (n = 736 of 55,120) provided essential data for 97.9% of all eligible cases. More specifically, we found 13 cases (n = 11 in the training set and n = 2 in the validation set) of 622 cases had no informative key terms in the retrieved clinical notes (Data Supplement); specifically, there were five cases without response, four cases without drug information, and four cases with neither response nor drug information. Various RECIST categories were applicable depending on specific therapeutic strategies (Table 1). For example, when patients received maintenance or consolidative pharmacotherapy, their RECIST responses would be SD or PD.

The resulting data set of 609 cases (n = 503 for training and n = 106 for validation purposes) and the two subgroups can be flexibly reassigned. The 736 fully annotated, deidentified clinical notes formed a data corpus with pharmacotherapies and the four response levels (CR, PR, SD, and PD). A basic description of demographic and clinical information of the data set cases is provided in Table 2. Of the total 609 patients, the mean age at the time of diagnosis was 64.3 years (range, 31.7-89.3 years); median follow-up time is 18.1 months (range, 1.2-242.4 months); non–small-cell lung cancer was the most frequent lung cancer subtype; and there is no significant difference in each variable between the two cohorts.

Detailed therapy regimens and specific drug combinations are available in the data set, as seen in Table 3, which illustrates the initial treatment drugs and combinations of the data set. Nearly 80% of the patients received doublet therapies, mainly with a platinum compound and another drug. Such information for subsequent treatment is also available (data not shown).

Working guidelines for searching and annotating key terms and attributes are provided in Table 4, which documented encountered alternative terms for "pharmacotherapy" in EHRs, which varied from "therapy" in the context of anticancer drug names, chemotherapy, specific drug names (eg, "etoposide and cisplatin," "paclitaxel," "tarceva," "gemcitabine"). Variations in four attributes are also documented. Data Supplement Table 7 summarizes Other conditions recorded in the

TABLE 3. Pharmacologic Drugs Used in Initial Systemic Therapy of the Gold Standard Data-Set Patients

Pharmacotherapy	Drug Names	Training Set (n = 503)	Validation Set (n = 106)	Total (N = 609)
Single agent	Individual drugs[a]	80 (15.9)	18 (17.0)	98 (16.1)
Doublets	Platinum[b] + etoposide	167 (33.2)	32 (30.2)	199 (32.6)
	Platinum + gemcitabine	15 (3.0)	3 (2.8)	18 (3.0)
	Platinum + taxanes[c]	160 (31.8)	37 (34.9)	197 (32.3)
	Platinum + other[d]	15 (3.0)	2 (1.9)	17 (2.8)
	Other doublets[d]	31 (6.2)	8 (7.6)	39 (6.4)
	Subtotal	388 (77.1)	82 (77.3)	470 (77.2)
Triplets or more	Platinum + taxanes + other(s)[d]	9 (1.8)	3 (2.8)	12 (2.0)
	Other three or more drugs[d]	26 (5.1)	3 (2.8)	29 (4.8)
	Subtotal	35 (7.0)	6 (5.7)	41 (6.7)

NOTE. Data reported as No. (%).

[a]Single drugs: ATM-1, bevacizumab, capecitabine, carboplatin, cediranib, cisplatin, docetaxel, erlotinib, etoposide, figitumumab, gefitinib, gemcitabine, irinotecan, cemadotin, mitomycin, paclitaxel, pemetrexed, sorafenib, temozolomide, temsirolimus, tipifarnib, topotecan, vinorelbine.

[b]Platinum: carboplatin, cisplatin, oxaliplatin.

[c]Taxanes: docetaxel, paclitaxel.

[d]Other: Bayer MMPI, bevacizumab, bleomycin, bortezomib, cyclophosphamide, doxorubicin, EKB 569, temsirolimus, etoposide, erlotinib, everolimus, figitumumab, fluorouracil, gefitinib, irinotecan, mitomycin, pemetrexed, rituximab, sorafenib, tanomastat, tanespimycin, tipifarnib, topotecan, vatalanib, vincristine, vinorelbine.

TABLE 4. Variation of Terms in the Corpus Defining Entities and Attributes of RECIST Categories

Complete Response	Partial Response	Stable Disease	Progressive Disease
Cancer free	Almost completely regressed	Approximately stable	Enlarged in size
CR	Almost completely regression	Clinically stable	Have a recurrence
Complete remission	Apparent response	Not changed	Mixed response
Complete response	Decrease in the size	No evidence of disease	In response to an increase
Complete resolution	Decreased tumor or adenopathy	No evidence of disease recurrence	Progression
Negative PET scan	Diminished in size	No major change	Progressive
No evidence of active disease	Dramatic decrease	No significant changes	Progression of disease
	Dramatic regression	Stable	Progressive cancer
	Dramatic response	Stable disease	Progressive disease
	Enough of a response	SD	PD
	Excellent response	Unchanged	Progressive enlargement
	Good response		Recurrent
	In response to treatment		Relapse
	Marked reduction		Significant progression
	Marked shrinkage		
	Much response		
	Near-complete regression		
	Near-complete response		
	Nice regression		
	Nice response		
	Partial regression		
	Partial remission		
	Partial response		
	Positive response		
	PR		
	PR response		
	Regression of disease		
	Remarkable response		
	Residual disease		
	Residual tumor or adenopathy		
	Respond favorably		
	Responding favorably		
	Respond nicely		
	Responding well		
	Shrinkage of the tumor		
	Shrink tumor markedly		
	Shrunk significantly		
	Significantly reduced in size		
	Significant response		
	Some response		

Abbreviations: CR, complete response; PD, progressive disease; PET, positron emission tomography; PR, partial response; SD, stable disease.

clinical notes of the data set cases as summarized in the Data Supplement; some are under more than one category because they could result from different causes.

The fully annotated data set will be released upon request to academic and noncommercial users. Although great efforts have been made to fully deidentify the data set, appropriate data use agreements will still be required.

DISCUSSION

We have created a gold standard data set of 609 cases, with built-in training and validation sets, to effectively capture all relevant information contributing to response levels as defined by RECIST. We demonstrated how to effectively reduce the seemingly insurmountable numbers of clinical notes: that is, 55,120 medical notes in 72 types, recorded in Mayo Clinic's EHRs, reduced to 736 that covered 609 patients—specifically, annotations of 1.3% of total clinical notes provided essential information for 97.9% of all eligible cases. Our objective was to support an expedited process of extracting antineoplastic therapy response information in a real-world setting of patients diagnosed with solid tumors, exemplified through lung cancer, with a specific goal to develop an NLP-supported RECIST tool for extracting and synthesizing antineoplastic therapy responses from EHRs in accelerating research into antineoplastic therapeutic response. Tools built from this resource could accelerate standardized evaluation of patient outcomes in a wide scope and is applicable to different drug groups and specific regimens of virtually all solid tumors when using RECIST.

Notably, there were several limitations in this study. One is that we only extracted the responses and all drug regimens combined. However, there were several confounding variables between the response and pharmacotherapy. For example, the relevance of response to multiple lines of chemotherapy, the various drugs, treatment of comorbidity and other conditions, surgical radicality, and adjuvant therapy, which caused complexity in the comparison. In our current study, we only selected the pharmacotherapy and response level as defined by RECIST. Because of heterogeneity of treatments and responses (eg, CR is unusual in metastatic non–small-cell lung cancer), the corpus has class imbalances that may limit generalizability to all clinical scenarios.

Another limitation is that we emphasized antineoplastic drugs, not including other treatment modalities. In real-world practice, there are diverse treatments for lung cancer, including surgery, drug therapy, radiation therapy, and palliative care. RECIST may be applicable to radiation therapy and palliative care interventions.

In terms of future directions and ongoing efforts, we are developing an NLP tool to accurately extract RECIST-based treatment responses in patients with cancer. Despite the widespread adoption of EHRs and multiple efforts to rapidly retrieve useful information to improve patient care, researchers remain challenged by the heterogeneity of EHRs data. Much of the information required to conduct precision medicine, which encompasses the individualized capture of responses to treatment, is contained within unstructured, written texts and clinical notes. In its current state, this information is not computable; hence, NLP offers an exciting opportunity to unlock these data. Menasalvas Ruiz et al[12] reported in 2018 on a first integration of an NLP framework for the analysis of clinical records of patients with lung cancer, making use of a telephone assistance service of a major Spanish hospital. The study specifically showed how relevant data (ie, patient demographics and medical comorbidities) can be extracted, and how these data can be used to conduct relevant analyses. Their study demonstrates that integration of unstructured EHR text within a data analysis framework is technically feasible and worthy of additional study. However, to our knowledge, there is no NLP system capable of extracting antineoplastic treatment response from EHRs. In fact, large commercial entities such as Flatiron Health (New York, NY) have specifically identified this as a highly challenging area.[7,13,14] The aim of our NLP tool is to quickly and easily enable response extraction for lung cancer pharmacotherapy in oncology and for additional medical questions, which can expand to specific antineoplastic drugs and extend to all solid tumors.

Our gold standard data set contains comprehensive information on each individual record that can be used for additional purposes beyond determining RECIST-defined responses. For instance, our framework can extract each line of treatment, patient comorbidity, toxicity/adverse event, or relapse/progression that occurs, as proposed in the following paragraphs.

Some treatment attributes were not retrieved due to responses being described imprecisely in clinical documentation. A goal is to link treatment responses to specific therapy regimens.

As depicted in Figure 1, treatment options for lung cancer include surgery, radiation therapy, chemotherapy, immunotherapy, and targeted therapy. Therapeutic-modality recommendations depend on several factors, including the type and stage of cancer. In this study, we only focused on pharmacotherapy and analyzed drug exposures regardless of stage. In future research, we will categorize drug exposures on the basis of treatment intent.

Treatment toxicities can vary significantly depending on the class of pharmacotherapy received. Distinguishing the toxicities associated with different classes of therapies on the basis of EHR data will be a challenging but crucial focus of future work.

In our current study, we only focused on evaluation of target lesions in the thorax. RECIST criteria can also be applied in evaluation of nontarget lesions in metastatic sites as well as pathologic lymph nodes.

Patients with lung cancer often have multiple comorbidities. The overall impact of preexisting conditions on lung cancer outcomes is not known and can be difficult to determine. Others have shown comorbidities predict overall survival independently in response to antineoplastic therapy, although the effects are often modest.[15,16] Extraction of comorbidity data captured within the EHRs and correlation of these data with long-term outcomes are a compelling focus of future work.

AFFILIATIONS

[1]Department of Health Sciences Research, Mayo Clinic, Scottsdale, AZ
[2]Division of Pulmonary & Critical Care Medicine, West China Hospital, Sichuan University, Chengdu, Sichuan, China
[3]Department of Chest Medicine, Taipei Veterans General Hospital, Taipei City, Taiwan
[4]Division of Biomedical Statistics and Informatics, Department of Health Science Research, Mayo Clinic, Rochester, MN
[5]Department of Medicine, Division of Hematology/Oncology, Vanderbilt University, Nashville, TN
[6]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX
[7]Department of Health Sciences Research, Mayo Clinic, Scottsdale, AZ
[8]Department of Biomedical Informatics, Vanderbilt University, Nashville, TN

CORRESPONDING AUTHOR

Ping Yang, MD, PhD, Division of Epidemiology, Department of Health Sciences Research, Mayo Clinic, 13400 E Shea Blvd, Scottsdale AZ; e-mail: yang.ping@mayo.edu.

SUPPORT

Supported by the National Institutes of Health (Grants No. CA77118, CA80127, CA84354 [P.Y.]; CA194215 [J.L.W.]; and HG008341 [H.X.]) and Mayo Foundation funds.

AUTHOR CONTRIBUTIONS

Conception and design: Yalun Li, Jeremy L. Warner, Hua Xu, Ping Yang
Provision of study material or patients: Jason A. Wampfler
Collection and assembly of data: Yalun Li, Yung-Hung Luo, Jason A. Wampfler, Ping Yang
Data analysis and interpretation: Yalun Li, Samuel M. Rubinstein, Firat Tiryaki, Kumar Ashok V, Ping Yang
Manuscript writing: All authors
Final approval of manuscript: All authors
Accountable for all aspects of the work: All authors

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated unless otherwise noted. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/cci/author-center.

Open Payments is a public database containing information reported by companies about payments made to US-licensed physicians (Open Payments).

Jeremy L. Warner
Stock and Other Ownership Interests: HemOnc.org
Consulting or Advisory Role: Westat, IBM
Travel, Accommodations, Expenses: IBM

Hua Xu
Employment: Melax Technologies
Stock and Other Ownership Interests: Melax Technologies
Consulting or Advisory Role: More Health, DCHealth Technologies, Hebta
Patents, Royalties, Other Intellectual Property: Royalties from software license from UTHealth

No other potential conflicts of interest were reported.

ACKNOWLEDGMENT

We thank Dr Hongfang Liu, PhD, and Liwei Wang, PhD, for their contributions at the design and research phases of this study.

REFERENCES

1. Wang Y, Wang L, Rastegar-Mojarad M, et al: Clinical information extraction applications: A literature review. J Biomed Inform 77:34-49, 2018
2. Wu S, Roberts K, Datta S, et al: Deep learning in clinical natural language processing: A methodical review. J Am Med Inform Assoc 27:457-470, 2020
3. Savova GK, Danciu I, Alamudun F, et al: Use of natural language processing to extract clinical cancer phenotypes from electronic medical records. Cancer Res 79:5463-5470, 2019
4. Eisenhauer EA, Therasse P, Bogaerts J, et al: New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). Eur J Cancer 45:228-247, 2009
5. Schwartz LH, Bogaerts J, Ford R, et al: Evaluation of lymph nodes with RECIST 1.1. Eur J Cancer 45:261-267, 2009
6. Dancey JE, Dodd LE, Ford R, et al: Recommendations for the assessment of progression in randomised cancer treatment trials. Eur J Cancer 45:281-289, 2009
7. Griffith SD, Tucker M, Bowser B, et al: Generating real-world tumor burden endpoints from electronic health record data: Comparison of RECIST, radiology-anchored, and clinician-anchored approaches for abstracting real-world progression in non-small cell lung cancer. Adv Ther 36:2122-2136, 2019
8. Yang P, Allen MS, Aubry MC, et al: Clinical features of 5,628 primary lung cancer patients: Experience at Mayo Clinic from 1997 to 2003. Chest 128:452-462, 2005
9. Sun Z, Aubry MC, Deschamps C, et al: Histologic grade is an independent prognostic factor for survival in non-small cell lung cancer: An analysis of 5018 hospital- and 712 population-based cases. J Thorac Cardiovasc Surg 131:1014-1020, 2006
10. Xie D, Allen MS, Marks R, et al: Nomogram prediction of overall survival for patients with non-small-cell lung cancer incorporating pretreatment peripheral blood markers. Eur J Cardiothorac Surg 53:1214-1222, 2018
11. Stubbs S: MAE and MAI: Lightweight annotation and adjudication tools. Proceedings of the 5th Linguistic Annotation Workshop 129-133, 2011
12. Menasalvas Ruiz E, Tuñas JM, Bermejo G, et al: Profiling lung cancer patients using electronic health records. J Med Syst 42:126, 2018
13. Griffith SD, Miksad RA, Calkins G, et al: Characterizing the feasibility and performance of real-world tumor progression end points and their association with overall survival in a large advanced non-small-cell lung cancer data set. JCO Clin Cancer Inform 3:1-13, 2019
14. Khozin S, Miksad RA, Adami J, et al: Real-world progression, treatment, and survival outcomes during rapid adoption of immunotherapy for advanced non-small cell lung cancer. Cancer 125:4019-4032, 2019
15. Iachina M, Jakobsen E, Møller H, et al: The effect of different comorbidities on survival of non-small cells lung cancer patients. Lung 193:291-297, 2015
16. Gould MK, Munoz-Plaza CE, Hahn EE, et al: Comorbidity profiles and their effect on treatment selection and survival among patients with lung cancer. Ann Am Thorac Soc 14:1571-1580, 2017