frontiers | Frontiers in Pharmacology

TYPE Hypothesis and Theory
PUBLISHED 15 September 2023
DOI 10.3389/fphar.2023.1180962

OPEN ACCESS

EDITED BY
Sheyu Li,
Sichuan University, China

REVIEWED BY
Koen Degeling,
The University of Melbourne, Australia
Federico Motta,
University of Modena and Reggio Emilia, Italy

*CORRESPONDENCE
Blythe Adamson,
badamson@flatiron.com

RECEIVED 06 March 2023
ACCEPTED 25 August 2023
PUBLISHED 15 September 2023

CITATION
Adamson B, Waskom M, Blarre A, Kelly J, Krismer K, Nemeth S, Gippetti J, Ritten J, Harrison K, Ho G, Linzmayer R, Bansal T, Wilkinson S, Amster G, Estola E, Benedum CM, Fidyk E, Estévez M, Shapiro W and Cohen AB (2023), Approach to machine learning for extraction of real-world data variables from electronic health records. *Front. Pharmacol.* 14:1180962. doi: 10.3389/fphar.2023.1180962

Approach to machine learning for extraction of real-world data variables from electronic health records

Blythe Adamson[1,2]*, Michael Waskom[1], Auriane Blarre[1], Jonathan Kelly[1], Konstantin Krismer[1], Sheila Nemeth[1], James Gippetti[1], John Ritten[1], Katherine Harrison[1], George Ho[1], Robin Linzmayer[1], Tarun Bansal[1], Samuel Wilkinson[1], Guy Amster[1], Evan Estola[1], Corey M. Benedum[1], Erin Fidyk[1], Melissa Estévez[1], Will Shapiro[1] and Aaron B. Cohen[1,3]

[1]Flatiron Health, Inc., New York, NY, United States, [2]The Comparative Health Outcomes, Policy, and Economics (CHOICE) Institute, Department of Pharmacy, University of Washington, Seattle, WA, United States, [3]Department of Medicine, NYU Grossman School of Medicine, New York, NY, United States

Background: As artificial intelligence (AI) continues to advance with breakthroughs in natural language processing (NLP) and machine learning (ML), such as the development of models like OpenAI's ChatGPT, new opportunities are emerging for efficient curation of electronic health records (EHR) into real-world data (RWD) for evidence generation in oncology. Our objective is to describe the research and development of industry methods to promote transparency and explainability.

Methods: We applied NLP with ML techniques to train, validate, and test the extraction of information from unstructured documents (e.g., clinician notes, radiology reports, lab reports, etc.) to output a set of structured variables required for RWD analysis. This research used a nationwide electronic health record (EHR)-derived database. Models were selected based on performance. Variables curated with an approach using ML extraction are those where the value is determined solely based on an ML model (i.e. not confirmed by abstraction), which identifies key information from visit notes and documents. These models do not predict future events or infer missing information.

Results: We developed an approach using NLP and ML for extraction of clinically meaningful information from unstructured EHR documents and found high performance of output variables compared with variables curated by manually abstracted data. These extraction methods resulted in research-ready variables including initial cancer diagnosis with date, advanced/metastatic diagnosis with date, disease stage, histology, smoking status, surgery status with date, biomarker test results with dates, and oral treatments with dates.

Abbreviations: AI, artificial intelligence; BERT, bidirectional encoder representations from transformers; EHR, electronic health records; LSTM, long term short memory; ML, machine learning; NPV, negative predictive value; NSCLC, non–small cell lung cancer; P&Ps, Policies and Procedures; PPV, positive predictive value; RWD, real-world data; RWE, real-world evidence.

Conclusion: NLP and ML enable the extraction of retrospective clinical data in EHR with speed and scalability to help researchers learn from the experience of every person with cancer.

KEYWORDS

electronic health records, cancer, oncology, real-world data, machine learning, natural language processing, artificial intelligence

Introduction

A barrier to generating robust real-world evidence (RWE) is access to research-ready datasets that demonstrate sufficient recency, clinical depth, provenance, completeness, representativeness and usability. Health outcomes must be appropriately defined and consistently measured. For studies using routinely collected electronic health record (EHR)-derived data, a considerable amount of data preprocessing and labor-intensive curation is required to create a dataset with clinically meaningful variables and outcomes needed for analysis (Figure 1).

The challenge is that so much valuable information is trapped within unstructured documents like clinician notes or scanned faxes of lab reports, where extracting the relevant data is far from trivial. The traditional approach to having clinical experts manually review patient charts to abstract data is time consuming and resource intensive (Birnbaum et al., 2020). This approach limits the number of patients available for research purposes. Learnings can quickly become outdated—for example as new biomarkers and treatments emerge, the standards of care change, or new indicators for social determinants of health are prioritized. In other instances, answers to important research questions remain infeasible due to limited sample sizes.

Artificial intelligence (AI) advances in the areas of natural language processing (NLP) and machine learning (ML) have created new opportunities to improve the scale, flexibility, and efficiency of curating high-quality real-world data (RWD) in oncology (Bhardwaj et al., 2017; Bera et al., 2019; Datta et al., 2019; Koleck et al., 2019; Shah et al., 2019; Wang et al., 2019; Bertsimas and Wiberg, 2020; Karimi et al., 2021; Subbiah, 2023). The definitions of foundational AI/ML terminology are provided in Table 1. When using ML and NLP for RWE, current guidance emphasizes transparency (NICE, 2022; Norgeot et al., 2020; Center for Drug Evaluation and Research Center for Biologics Evaluation and Research Oncology Center of Excellence; Padula et al., 2022; Blueprint for trustworthy AI implementation guidance and assurance for healthcare, 2022). The United Kingdom National Institute for Health and Care Excellence instructs that "where human abstraction or artificial intelligence tools are used to construct variables from unstructured data, the methods and processes used should be clearly described." (NICE, 2022).

In response to guidance, the objective of this paper is to describe the general approach for applied NLP and ML methods that are used by Flatiron Health to extract data from unstructured documents stored in oncology care EHR. A key distinction in our terminology is the use of "abstraction" meaning performed by humans and "extraction" meaning performed by models. Out of scope for this paper are other AI, ML, and NLP innovations and contributions from Flatiron Health, such as: model-assisted cohort selection (Birnbaum et al., 2019; Birnbaum et al., 2020); continuous bias monitoring software (Birnbaum et al., 2023); automated mapping of laboratory data (Kelly et al., 2022); prediction of future health events (Chen et al., 2019); and point-of-care products to improve patient care and clinical trials (Lakhanpal et al., 2021; Coombs et al., 2022).

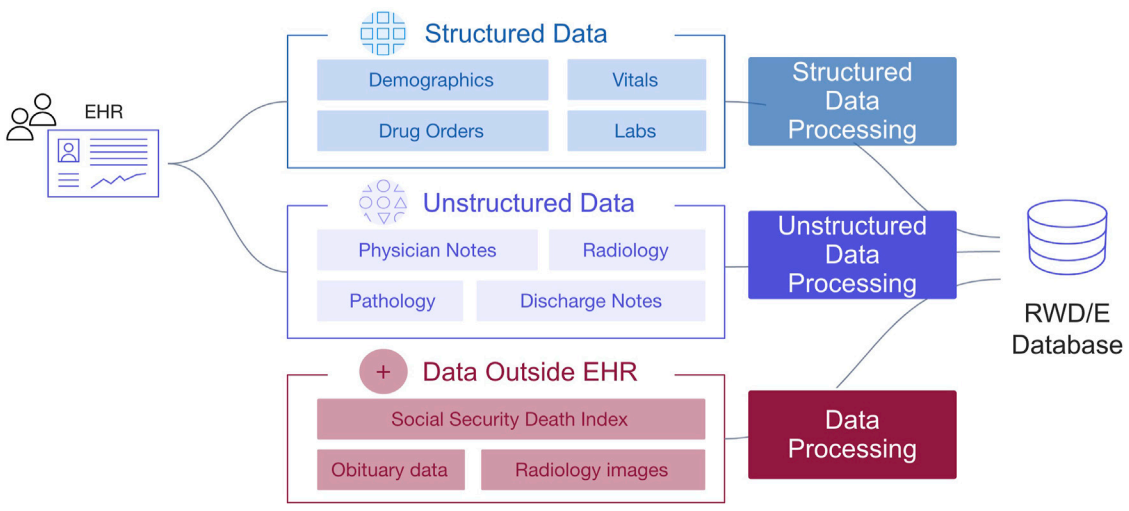

FIGURE 1
Overview of data variables defined by structured and unstructured information in EHR.

TABLE 1 Key terms in machine learning.

Foundational machine learning (ML) definitions
• **Class:** One of the possible values that a binary or categorical variable can take
• **Labels:** The known classes associated with data used to train or evaluate an ML model
• **ML-Extracted:** Algorithmic extraction of data from documented evidence in the patient chart (either structured or unstructured) at the time of running the model. Techniques include ML and natural language processing (NLP), in contrast to other data processing methods such as abstraction or derivation
• **Model:** An ML algorithm with a specific architecture and learned parameters that takes inputs (e.g., text) and produces outputs (e.g., extracted diagnosis)
• **NLP:** A field of computational systems (including but not limited to ML algorithms) that enable computers to analyze, understand, derive meaning from, and make use of human language
• **Score:** A continuous output from a model that can be interpreted as the model-assigned probability that a data point belongs to a specific class
• **Threshold:** A cutoff value that defines classes when applied to continuous scores. Binary variables (e.g., whether a patient has had surgery) have a natural default threshold of 0.5, but different thresholds might be leveraged depending on the relative tolerance for false positives vs false negatives required
Performance metric definitions
• **Sensitivity (Recall):** The proportion of patients abstracted as having a value of a variable (e.g., group stage = IV) that are also ML-extracted as having the same value
• **Positive predictive value (PPV) (Precision):** The proportion of patients ML-extracted as having a value of a variable (e.g., group stage = IV) that are also human abstracted as having the same value
• **Specificity:** The proportion of patients abstracted as not having a value of a variable (e.g., group stage does not = IV) that are also ML-extracted as not having the same value
• **Negative predictive value (NPV):** The proportion of patients ML-extracted as not having a value of a variable (e.g., group stage does not = IV) that are also abstracted as not having the same value
• **Accuracy:** The proportion of patients where the ML-extracted and abstracted values are identical. For variables with more than 2 unique values (e.g., group stage), accuracy within each class is calculated by binarizing the predictions (e.g., for Accuracy of group_stage = IV, all abstracted and ML-extracted values would be defined as either "IV" or "not IV"
• **F1 Score:** Computed as the harmonic mean of sensitivity and PPV. For a binary classifier, the threshold that maximizes F1 can be considered the optimal balance of sensitivity and PPV.

Materials and methods

Overview

We developed a set of research analysis variables using information from the documents available in patient charts. Variables were selected for exploration of ML extraction if commonly required for retrospective observational studies in oncology, but not consistently available in claims data or structured EHR data, and high-quality training data were available that had been manually curated by experts to produce a large amount of abstracted data available for training models (Haimson et al.).

The variables curated through our ML extraction approach are those where the values are solely derived from the identification of clinical details in the EHR documents by an ML model in combination of NLP techniques and rules-based logic. It is important to note that these values are not predictions or inferences, but rather a direct extraction of information that is clearly documented in the EHR.

EHR-derived data source

This study used the nationwide Flatiron Health EHR-derived de-identified database. The Flatiron Health database is a longitudinal database, comprising de-identified patient-level structured and unstructured data (Birnbaum et al., 2020; Ma et al., 2023). At the time of this research, the database included de-identified data from approximately 280 US cancer practices (~800 distinct sites of care).

Structured and unstructured data modalities are available in the database. EHR structured data elements include, but are not limited to, documented demographics (e.g., year of birth, sex, race/ethnicity, etc.), vitals (e.g., height, weight, temperature, etc.), visits, labs, practice information, diagnosis codes, medication orders, medication administrations, ECOG performance status, health insurance coverage, and telemedicine (Figure 1). EHR unstructured data and documents include, but are not limited to, paragraphs of clinic visit notes, PDF scans of lab results, radiology images with reports, pathology reports, and communications between the patient and care team (Figure 2). For the purpose of this paper, all the figures contain fictional representations of documents, sentences, dates and patient IDs.

Patient population

The large general cross-tumor cohort includes all patients with at least one International Classification of Diseases (ICD)-9 or ICD-10 cancer code and at least one unique-date clinic encounter

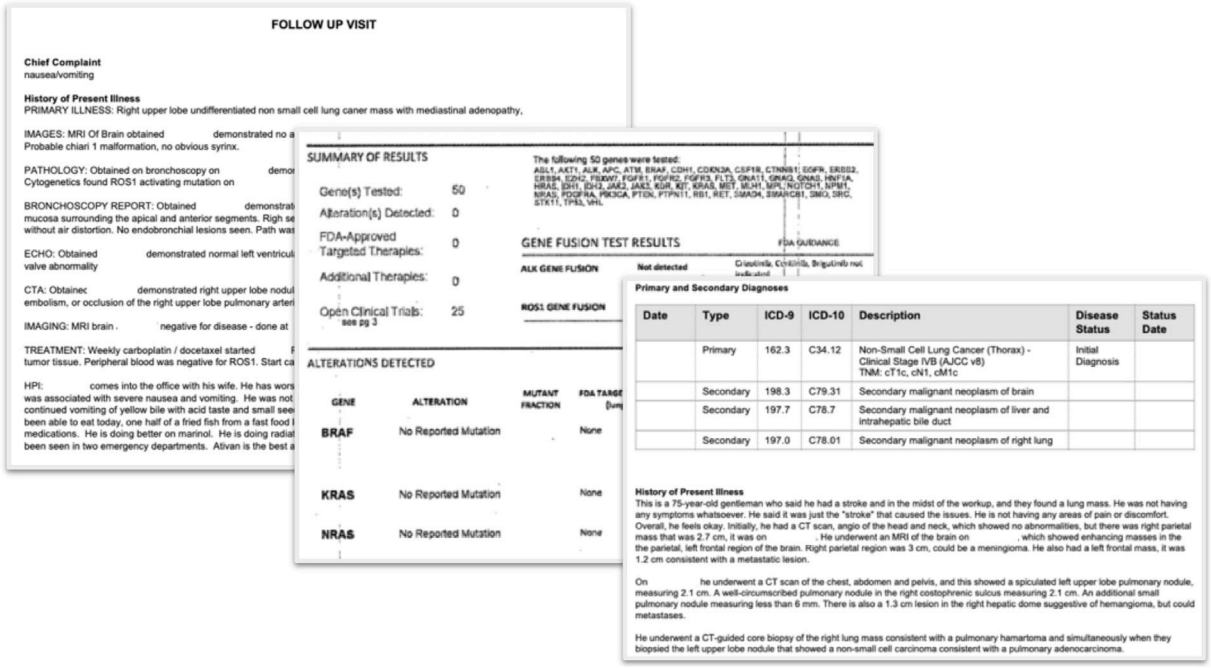

FIGURE 2
Examples of unstructured documents from EHR that are used as inputs for ML-extraction of information (all dates and patient IDs are fictitious).

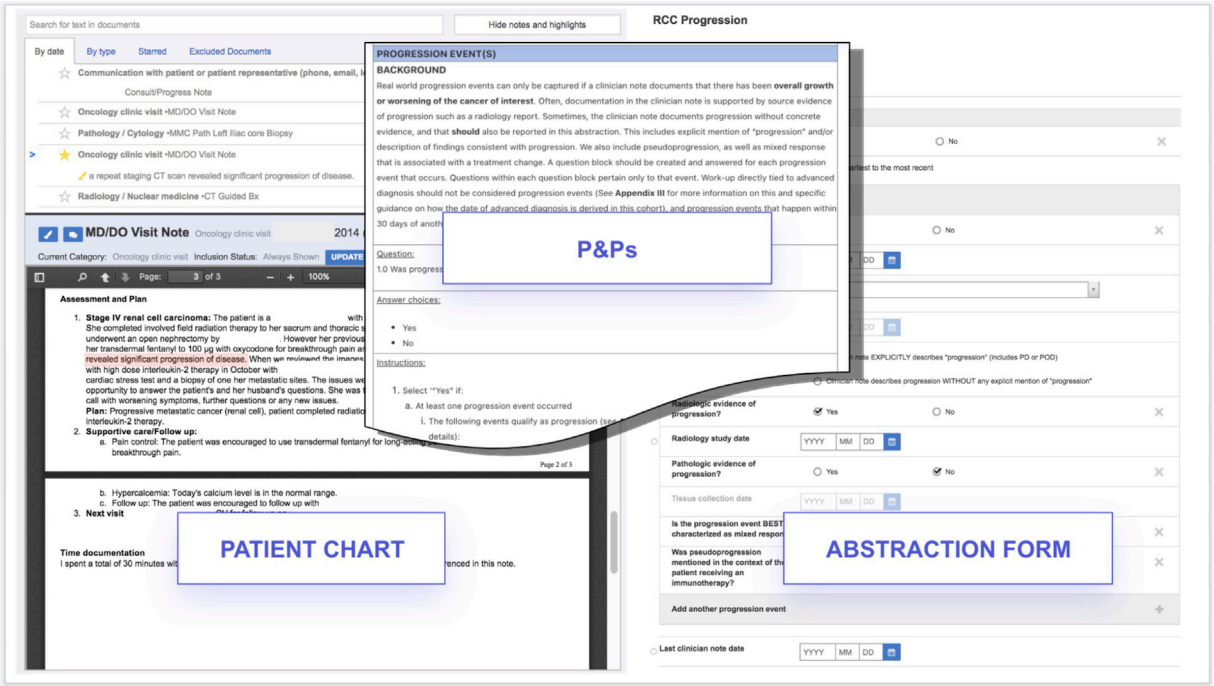

FIGURE 3
Technology enabled expert abstraction. Abbreviations: P&Ps, Policies and Procedures. All dates and patient IDs are fictitious.

documented in the EHR (reflected by records of vital signs, treatment administration, and/or laboratory tests) on or after 1 January 2011. 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 can vary between cancer types.

Clinical expert abstraction of variables for model development

Critical information in patient charts has been manually abstracted by trained clinical experts (i.e., clinical oncology nurses or tumor registrars), following a set of standardized policies and procedures. To abstract data from patient charts, we use a foundational technology (Shklarski et al., 2020) that enables clinical experts to more easily review hundreds of pages of documents to determine patient characteristics, treatments, and outcomes documented in the EHR (Figure 3).

Years of manual abstraction by a workforce of thousands of abstractors at Flatiron Health have created a large and high-quality corpus of labeled oncology EHR data. Clinically-relevant details specific to each cancer type are abstracted from every form of clinical documentation available in the EHR, including clinic visit notes, radiology reports, pathology reports, etc. Abstractors are trained to locate and document relevant information by following policies and procedures tested and optimized for reliability and reproducibility through iterative processes, and oversight is provided by medical oncologists.

The abstraction process undergoes continuous auditing to monitor abstractor performance, while proprietary technology links each curated data point to its source documentation within the EHR, enabling subsequent review. 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.

Flatiron Health has abstracted sets of clinically meaningful variables from more than 300,000 people with cancer to develop disease-specific de-identified research-ready databases (Ma et al., 2023). Limited by the capacity of human abstractors, there had remained millions of patients with cancer in the Flatiron Health database for whom no unstructured data had yet been curated to create variables with the clinical depth needed to generate meaningful insights. If a hypothetical variable required 30 min of chart review by a clinical expert to abstract the information of interest for 1 patient, then it would take a team of 100 full-time abstractors more than 7 years to finish defining 1 variable for a population of 3 million patients.

Overview of machine learning extraction approach

The objective of this application of NLP and ML methods was to replicate the expert abstraction process described in the previous section. When developing ML models for extracting information, all of the clinical abstractor expertise that was incorporated into the manual abstraction of variables is available to learn from through training. Once iterated upon and placed in production, ML models can automate information extraction from unstructured clinical data sources in a way that mimics expert clinical abstractors. The models expand on previously established technology infrastructure that includes deep learning architectures (Rich et al., 2023), text snippet-based modeling approaches (Birnbaum and Ambwani), and extraction of patient events and dates (Gippetti et al.; Ballre et al., 2022; Rich et al., 2022).

Alongside the manually-abstracted labels, we use NLP to pull relevant textual information from charts to use as inputs to train built-for-purpose ML models and model architectures for a given extraction task. Through this process we can make our end variables appropriate for disease-specific or pan-tumor (i.e., histology-independent) applications. For example, by deciding whether or not to use model training data sourced from curated disease-specific cohorts or any-cancer cohorts, we can make our model's output variables built-for-purpose to be used in an analysis that generates meaningful RWE for a specific research question.

A range of model architectures were evaluated and considered for the purpose of information extraction for variables of interest. The model output of variable classes ranged, including:

- binary (e.g., metastatic diagnosis Yes/No)
- categorical unordered (e.g., never smoker, history of smoking, current smoker)
- categorical ordered (e.g., cancer stage I-IV)
- date (e.g., 02/05/2019 start of oral treatment X)

Date and classification can come from the same model, separate models, or connected models.

Natural language processing to generate model inputs

For each variable of interest, we begin with clinical experts constructing a list of clinical terms and phrases related to the variable. Since models are trying to extract explicit information from charts, rather than infer it, only terms that are directly relevant to a specific variable are included (e.g., when extracting a patient's histology, terms could include "histology," "squamous," and/or "adenocarcinoma," but do not include treatment or testing terms from which the histology might be indirectly inferred).

Next, we use NLP techniques 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. The approach uses optical character recognition (OCR) systems to extract text from PDFs, faxes, or scans containing images; the text is then searched for relevant clinical terms. The contextual information surrounding the clinical term is critical because the words at the beginning of a sentence may change the interpretation of a key word at the end of a sentence. ML models can understand if the clinical concept appears and under what context—such as, if negativity, speculation, or affirmation exists in the surrounding clinical terms (i.e., snippets). Where applicable, any associated dates within these sentences are also identified. These sentences are then transformed into a mathematical representation that the model can interpret. The output of this document processing is a broad set of features aimed at fully capturing document structure, chronology, and clinical terms or phrases.

Machine learning model development

Features and labels

The features defined by NLP become the inputs provided to the model to score the likelihood that a given patient document is

associated with each class of a particular categorical variable (e.g., histology categories of non–squamous cell carcinoma, squamous cell carcinoma, non–small cell lung cancer [NSCLC] histology not otherwise specified). The final model output is the variable value for each patient. The labeled dataset is commonly split into three subsets: a training set, a validation set, and a test set. The training and validation sets are used to build the model, which often involves an iterative development process, while the test set is used to evaluate the performance of the final ML model.

Model development

The training set comprises labeled data points that are used to optimize the model's parameter values. In an iterative process, training examples are provided to the model, its outputs are compared to the labels, and the parameter values are adjusted in response to errors. By using manually-abstracted values as labels, the objective of this process is for the model to learn what answer a human abstractor would give when reading a specific clinical text.

The validation set is used to assess how well the model has learned these associations. Because the model does not see any data from patients in the validation examples during training, they can be used to estimate how it will perform on new, unlabeled examples once it is put into production. Validation performance is commonly assessed using metrics such as precision, recall, and F1 score (See Table 1 Key Terms in Machine Learning). These aggregate metrics, combined with review of individual errors, inform decisions about search terms, text preprocessing steps, and model architectures. Experimentation continues until a final "best" model is identified.

When a ML model is trained to perform a classification task, it outputs scores for each possible class for each data point. These scores are between 0 and 1 and show the probability that a patient belongs to each class, based on information in their electronic health record. However, the scores may vary if the wording in the records is unusual or if there is conflicting information. For example, if a patient's cancer stage is being restaged, there may be multiple mentions of different stages in the record, and the model may assign moderate scores to each stage if the restaging event is unclear.

To produce a discrete class value, the class with the highest score is often chosen, but other approaches may optimize performance. In particular, a probability threshold may be chosen such that a patient will be classified into one class if and only if their score exceeds the threshold. The optimal threshold depends on factors such as class balance and is typically chosen empirically (Lipton et al., 2014). When no class receives a sufficiently high score, another option is to defer to abstraction to resolve uncertainty (Waskom et al., 2023).

We explored and experimented with a range of ML models and architectures for the purpose of extracting specific variable information from the EHR. Deep learning architectures included long short-term memory (LSTM), Gated recurrent units (GRU), and bidirectional encoder representations from transformers (BERT) (Hochreiter and Schmidhuber, 1997; Shickel et al., 2018; Devlin et al., 2018). These models can learn thousands or millions of parameters, which enable them to capture subtleties in the text. They read sentences as a whole and use the words around a clinical term to incorporate surrounding context when determining the extracted class. When they receive very large texts as inputs, they can figure out where the relevant information is and focus on this section and its context.

For example, in LSTMs, words are passed into the model sequentially; during each step through a sentence, the model has access to memory (i.e., an internal state) that is impacted by the previous word, in effect allowing the model to "remember" the previous word (Figure 4). The LSTM block combines the new word with the information that came before to derive a more contextually rich representation of the word. For instance, when the LSTM reads the word "Advanced," it remembers (via the model's internal state) that it was preceded by the word "not" and is more likely to classify the patient as "not advanced."

Model evaluation and performance assessment

Once iteration on the ML model is complete, final model performance is measured on a test set that uses manually-abstracted labels as the source of truth. Test sets are designed to be large enough to power both top-level metrics and sub-group stratifications on a "held out" set, that is, on data not used to train the ML model or validate performance during prototyping. This allows the test set to assess the model's ability to correctly classify data points that it has never seen before, which is typically referred to as the "generalization" of the model.

Measuring performance is a complex challenge because even a model with good overall performance might systematically underperform on a particular subcohort of interest, and because while conventional metrics apply to individual models, dozens of ML extracted variables may be combined to answer a specific research question. We use a research-centric evaluation framework (Estévez et al., 2022) to assess the quality of variables curated with ML. Evaluations include one or more of the following strategies: 1) overall performance assessment, 2) stratified performance assessment, and 3) quantitative error analysis, and 4) replication analysis. As variables curated with NLP and ML are expected to be incorporated into the evidence generated that will guide downstream decision-making, variable evaluation can also include replication of analyses originally performed using abstracted data. Replication analyses allow us to determine whether ML-extracted data—either individual variables or entire datasets—are fit-for-purpose in specific use cases by assessing whether they would lead to similar conclusions.

Specific variable-level performance metrics are only interpretable for cohorts with characteristics that are similar to the test set, depending on inclusion criteria such as the type and stage of cancer. As a result, we do not report them here.

Python was the primary coding language used in the development of ML models described here. Institutional Review Board approval of the study protocol was obtained before study conduct, and included a waiver of informed consent.

Results

We successfully extracted key information from unstructured documents in the EHR using the developed proprietary ML models trained on large quantities of data labeled by expert abstractors. For this paper, we are focusing the results on examples within NSCLC as

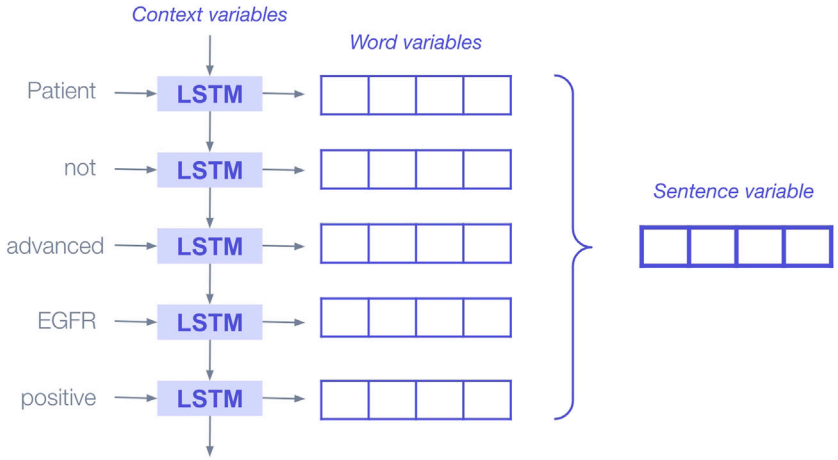

FIGURE 4
Illustration of deep learning bidirectional LSTM blocks applied sequentially to produce representations (aka, embeddings or encodings) that encapsulate the information added to the sentence by each new word. Abbreviations: LSTM, long short-term memory.

they were the first applications we developed. A set of 10 ML models output 20 distinct RWD variables for analysis, including initial cancer diagnosis with date, advanced/metastatic diagnosis with date, disease stage, histology, smoking status, surgery details, biomarker test results, and oral treatments with dates. Language snippets were the inputs for these models to produce a data point for each patient for each variable as outputs, illustrated in Figure 5.

Datatables containing variables curated by an approach using ML had the same appearance and functionality as variables curated with an approach using technology-enabled expert human abstraction (Figure 6).

Models had high performance when trained for disease-specific applications as well as histology-independent (i.e., tumor agnostic) patient cohorts. For example, the NSCLC specific *Histology Type* model had a sensitivity of 96% and a PPV of 94% for extracting non–squamous histology for patients with NSCLC. Detailed performance metrics are out of scope for this paper. Beyond satisfactory ML metrics, we found that in some cases ML-extraction can achieve similar error rates as manual abstraction by clinical experts (Waskom et al., 2023), and replication studies suggest that research analysis relying on multiple variables can reach similar results and conclusions when using variables curated by ML-extraction compared with human experts (Benedum et al., 2022; Sondhi et al., 2022; Benedum et al., 2023).

Approaches and learnings related to specific variables are described below.

Application examples

We have developed ML models for a number of different variables and use cases. A few of the more prominent models and their associated use cases are described below.

Cancer diagnosis and dates

We successfully developed deep learning models focused on the task of extracting initial, advanced, and metastatic cancer diagnosis and the corresponding diagnosis dates. Historically, ICD codes have been used as a proxy for diagnosis, as they are well captured in structured EHR data due to their use in billing. However, we have seen that the precision of ICD codes varies by disease, is not strongly correlated with disease prevalence in the larger population, and can be lower than 50%. With that in mind, extracting accurate diagnosis information is imperative to understanding patient populations, as errors at the diagnosis level propagate to all other variables. These models build on prior foundational research on extracting information from longitudinal clinic notes (Zhao et al., 2021; Agrawal et al., 2018). The initial, advanced, and metastatic variables are generated using multiple, distinct ML models. A conceptual diagram of this approach used by the metastatic variable is presented in Figure 7. We have found success chaining the models together—providing the output of one model as the input to the next—to prevent conflicting predictions and improve overall accuracy. An early investigation into model performance has been presented previously (Rich et al., 2021).

Additional complexity exists when trying to identify patients with rare cancers, primarily due to the low number of labels. We have demonstrated that techniques such as generic token replacement and leave-one-out validation can be effective in combating these complexities—allowing our models to successfully generalize to rare diseases, with few or no labels provided during training from the target disease(s).

Disease stage and histology

We successfully developed a deep learning model to extract cancer stage information and a second ML model to extract the histology of the tumor. One example of how we used this approach for a disease-specific application was training on patients with NSCLC. This model was designed to extract main stage (I-IV) and substage (letters A-C) granularity. Histology was extracted as a non-ordered categorical variable with the possible variable values of non–squamous cell carcinoma, squamous cell carcinoma, or NSCLC histology not otherwise specified.

Deep Learning Model Name	Language in Source EHR as Illustrative Snippet (Model Input)	Extracted Variables (Model Output)
Initial Diagnosis	"Mr. Smith was initially diagnosed with stage IIa NSCLC on 03-31-2017"	*IsDisease, DiseaseDiagnosisDate*
Advanced Diagnosis	"Unfortunately, Mr. Smith developed recurrence of his NSCLC on 09-01-2018"	*IsAdvanced, AdvancedDiagnosisDate*
Metastatic Diagnosis	"Mr. Smith was diagnosed with metastatic NSCLC on 03-31-2017"	*IsMetastatic, MetastaticDiagnosisDate*
Stage	"Mr. Smith was diagnosed with Stage IV NSCLC on 03-31-2017"	*GroupStage*
Histology	"Mr. Smith's biopsy showed a diagnosis of adenocarcinoma of the lung."	*HistologyType*
Smoking Status	"Mr. Smith is an 80-year non-smoker…"	*SmokingStatus*
Surgery	"Mr. Smith underwent wedge resection of his biopsy proven NSCLC."	*HasSurgery, SurgeryDate*
Biomarkers	"Mr. Smith received NGS test results on 2/15/2020 for EGFR, ALK, and ROS1 and was found to have an ALK rearrangement."	*BiomarkerName, BiomarkerStatus, ResultReturnedDate*
PD-L1	"Mr. Smith was diagnosed with adenocarcinoma of the lung, PD-L1 <1% on 2/20/2021."	*BiomarkerName, BiomarkerStatus, ResultReturnedDate*
Orals	"...she has received erlotinib since May 15th 2017 but stopped on Sept 15th 2017 for progression. She was then started on osimertinib on 9/25/2017 and remains on it currently."	*DrugName, StartDate, EndDate*

FIGURE 5
Sentences (fictional examples here) from EHR are inputs to deep learning models that produce a data variable value for each patient as an output. Language snippets are only extracted around key terms from which a variable might be extracted, and not around terms from which it could be indirectly inferred. Abbreviations: EHR, electronic health record; PD-L1, programmed death ligand 1. All dates and patient IDs are fictitious.

As cancer stage is documented similarly across solid tumor diseases, we were able to scale our approach to extract disease stage in a tumor-agnostic cohort with a similar deep learning architecture but training data composed of patients with multiple cancer types. While hematologic cancers have some important differences from solid organ cancers when it comes to assigning stage (risk stratification scores, no concept of metastatic disease, etc.), we found success using a deep learning model to extract this information for a number of hematologic cancers. Tumor histology is not as straightforward to scale across cancer types, as different cancers originate from different possible cell types (and therefore have different histologies). This means that to date, we use distinct histology models for each type of cancer. Performance evaluations for disease stage and histology are conducted at each category level and by cancer type as appropriate for use cases.

Smoking status

We successfully developed a deep learning model to extract information in the patient chart that indicates whether or not the patient has any lifetime history of smoking. The categorical variable output has the possible values as history of smoking, no history of smoking, or unknown. The most relevant sentences for this model were most often found in social history paragraphs of text that is a standard section in clinical encounter notes. Critical document categories that enabled high accuracy of this model included access to oncology clinic visit notes, radiology reports, surgery reports, lab reports, and pulmonary test result reports. The smoking status model was trained on a broad dataset of patients that included many cancer types for whom we have abstracted smoking status.

Surgery and surgery date

We successfully developed a deep learning model to extract information about whether the patient had a primary surgical procedure where the intent was to resect the primary tumor. As these types of surgeries often happen in outpatient facilities or hospitals, this valuable documentation lives in unstructured text formats in the oncology EHR. We have abstracted surgery data in certain disease cohorts but, because of the similarity in documentation approaches across cancer types, we were able to

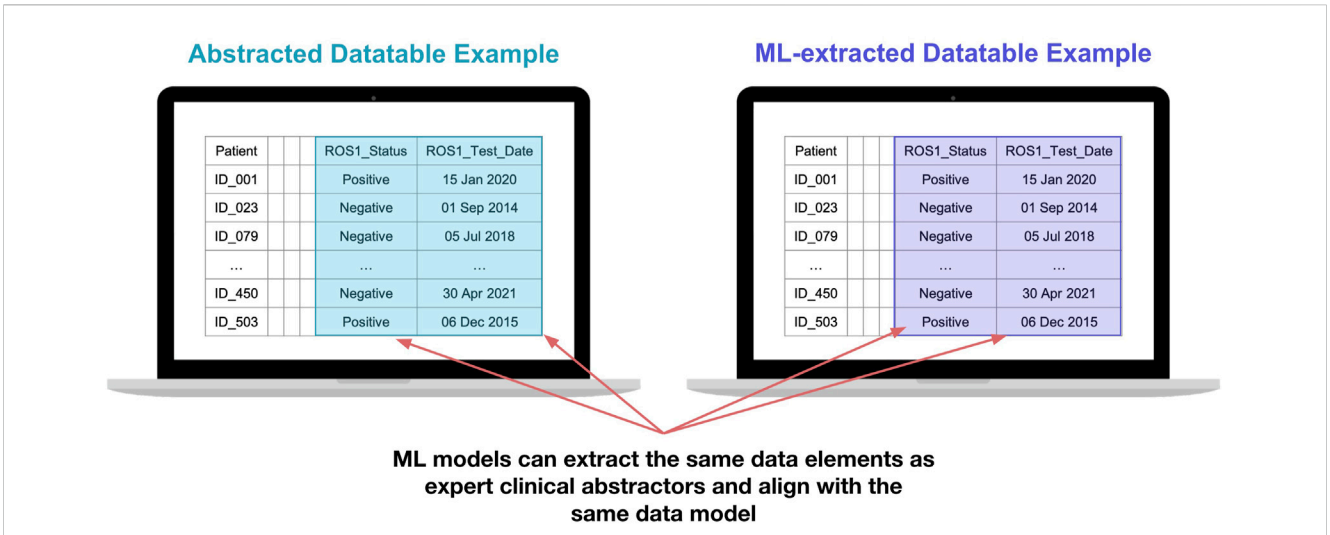

FIGURE 6
Illustration of a data table with variables curated by an approach using expert human abstractors (right) alongside a data table with variables curated by an approach using deep learning models (left) shows opportunity for exchangeable utility in real-world data analysis. All dates and patient IDs are fictitious.

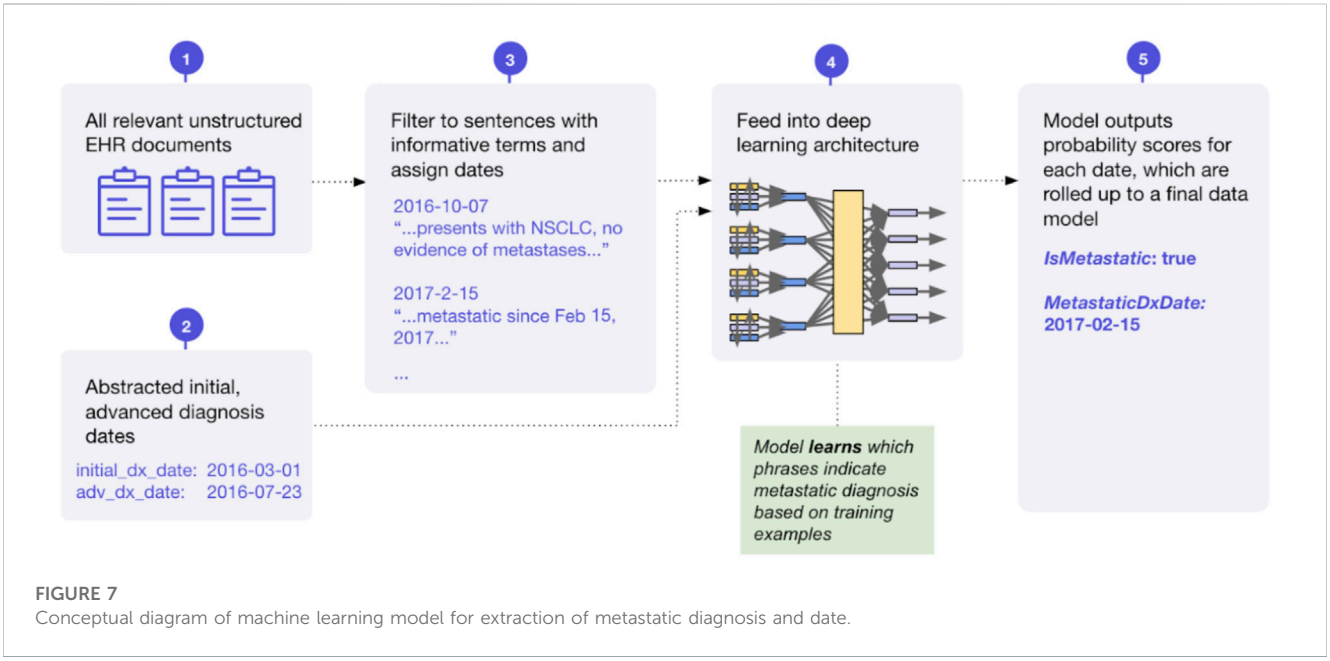

FIGURE 7
Conceptual diagram of machine learning model for extraction of metastatic diagnosis and date.

train a model that is tumor agnostic. This allowed us to scale surgery status and date in larger patient populations and in new disease types.

Biomarker testing results and result date

We successfully developed and deployed models to generate variables for biomarker testing, including extraction of the dates that the patient had results returned (Figure 6). One part of the model is able to identify whether or not a given document for a patient contains a biomarker test result. A separate part of the model is able to extract from the document the date a result was returned and the biomarker result. Early efforts with a regularized logistic regression model were presented previously (Ambwani et al., 2019) and more sophisticated models have been developed since.

A model first cycles through every EHR document for a given patient to understand whether or not the document contains biomarker testing results. These models rely on access to lab reports, including those saved in the EHR as a PDF or image of a scanned fax. The models are able to process report documents produced by different labs (e.g., Foundation Medicine, Caris, Tempus, etc.,) in addition to the clinician interpretations in visit notes.

A separate model then extracts the biomarker (e.g., included but not limited to *ALK, BRAF, EGFR, KRAS, MET, NTRK, RET, ROS1*,

or PD-L1) and test result (e.g., positive, negative, or unknown). This approach gives our ML models flexibility to extract biomarkers that the model may not have seen before in training. For PD-L1, where results are quantitatively reported, a separate ML model was developed to extract percent staining, with classes of <1%, 1%–49%, ≥49%, and unknown.

Since patients can receive biomarker testing multiple times throughout the treatment journey and at multiple facilities, it is possible that a given patient has more than one biomarker test result and date. For each patient, this allows us to determine biomarker status at different clinical milestones (e.g., advanced diagnosis date, start of second-line treatment, etc).

Oral treatments and treatment dates

We successfully developed a deep learning model to extract oral treatment information, including the treatment name, and the span for which the treatment was administered. In contrast to intravenous therapies such as chemotherapy or immunotherapy in which each dose is ordered and administered to be given in the clinic or infusion room, oral therapies are prescribed to patients to be filled by an outpatient pharmacy, which is frequently outside the clinic site. To have a complete understanding of all cancer treatments received or delayed (e.g., postponed during a hospitalization), it is necessary to enumerate the use of oral treatments through review of unstructured clinician visit notes, prescriptions, and communications with the patient or patient representative. Important information to select within the paragraphs of text include the treatment name, start date, and end date. We previously published an initial framework (Agrawal et al., 2018) for extracting drug intervals from longitudinal clinic notes, prescriptions, and patient communication documents and have developed more sophisticated and accurate methods since then. We found the visit notes contained key pieces of information about treatments being held or started when patients were hospitalized.

The model is trained to select mentions of a specific list of drug names used for oral treatment in the specific cancer type, along with the start date and end date. These oral treatment variables are generated using three distinct ML models. The list of oral treatments of interest were specific to each disease and defined by oncology clinicians. Expert abstraction from charts includes policies and procedures for collection of treatment start dates and discontinuation dates as both are needed to execute many common RWE study designs. To be fit for purpose, ML models were trained to extract both start and end dates of treatments.

Discussion

This paper described one approach to curating real-world oncology data variables from unstructured information in EHR using NLP and ML methods. Model development was possible with access to a large and high-quality corpus of labeled oncology EHR data produced via manual abstraction by a workforce of thousands of clinical expert abstractors over the course of several years. We now have models that are able to meet or even exceed human abstraction performance on certain tasks (Waskom et al., 2023). Using a performance evaluation framework (Devlin et al., 2018) for variables curated using the approach of ML extraction we affirmed high quality and fitness-for-use in RWE generation. We have shown that validations using the combination of multiple ML-extracted variables in one RWD analysis demonstrated no meaningful difference in RWE findings based on replications with the Flatiron Health variables curated by ML extraction compared with expert human abstraction (Forsyth et al., 2018; Zeng et al., 2018; Jorge et al., 2019; Karimi et al., 2021; Maarseveen et al., 2021; Benedum et al., 2022; Sondhi et al., 2022; Yang et al., 2022; Benedum et al., 2023).

Crucial information about clinical details may be recorded only within free-text notes or summaries in unstructured EHR documents. Our models primarily rely on deep learning architectures, because curating data from such sources usually requires techniques that capture the nuances of natural language. We select model architectures on a case-by-case basis depending on what works best for each variable, but we have found that the quality of the training data and labels can be just as, if not more, important to success than the architecture used. Despite this, we do expect that advances in generative AI and advancing LLM architectures will make deeper and more nuanced clinical concepts accessible to ML extraction, as LLMs are able to take into account a fuller context of the patient data and rely less on having high quality labels for training. The impressive generative abilities of models like gpt3 and its ChatGPT application have demonstrated this, although the generative framework itself may remain more suited for tasks such as summarization (Adams et al., 2021) than for scalable curation of structured real-world datasets.

The mission to improve and extend lives by learning from the experience of every person with cancer is more important than ever. With increasingly specific combinations of patient characteristics, disease, and therapy, we need to learn from as many relevant examples as possible to have statistically meaningful results. ML expands the opportunity to learn from patients who have been oppressed or historically marginalized in oncology clinical trials (Adamson et al., 2019; Hooley et al., 2019). As oncology care rapidly evolves, and the treatment landscape becomes more personalized—targeting new biomarkers, finely tuned to increasingly particular patient profiles—transparent fit-for-purpose applications of ML will have increasing importance. This will be valuable to gain trust with decision-makers in applications such as postmarket safety surveillance. With high performance models, we can truly learn from every patient, not just a sample. It also creates an opportunity to improve the completeness of RWD variables that were previously defined by only structured data elements, reducing potential bias in evidence.

There are strengths and limitations to the EHR curation approaches described here. Strengths include the large size, representativeness, and quality of training data used; success across a multitude of cancer types; and the explainability of approach to finding clinical details in documents. Massive volumes of high-quality expert abstracted data were a unique advantage for training high-quality ML models. Researchers at Stanford have confirmed similar capabilities with a different EHR dataset—detecting the timeline of metastatic recurrence of breast cancer (Banerjee et al., 2019). An example of a variable that would be challenging for ML extraction could be

microsatellite instability (MSI), where results are reported in a wide range of formats. One of the formats is a graphic where the result is reported visually on a sliding scale rather than in text format. This would be difficult for a model that relies on interpretation of text. The ML models described here were trained for and applied only in a US population (Ma et al., 2023). While the most suitable model architectures for each variable may be transferable across country borders, a limitation of this approach is that models must be re-trained with local data for highest performance.

The capability to build ML models that can extract RWD variables accurately for a large number of patients further enables the possible breadth and depth of timely evidence generation to answer key policy questions and understand the effects of new treatment on health outcomes.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: 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.

Author contributions

BA, JK, SN, and EE contributed to the conception of this review paper. AB, GH, GA, JK, JG, JR, KH, KK, MW, RL, TB, and SW developed the ML models. CB, ME, EF, AC, and BA conducted performance evaluations and validations. BA wrote the first draft of the manuscript. SN, GA, WS, MW, ME, AB, AC, EF, and RL wrote sections of the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This study received funding from Flatiron Health. The funder was involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

Acknowledgments

The authors would like to thank Flatiron Health's Selen Bozkurt, Sharang Phadke, Shreyas Lakhtakia, Nick Altieri, Qianyu Yuan, Geetu Ambwani, Lauren Dyson, Chengsheng Jiang, Somnath Sarkar, Javier Jimenez, Arjun Sondhi, Alexander Rich, Benjamin Birnbaum, Andrej Rosic, Barry Leybovich, Jamie Irvine, Nisha Singh, Sankeerth Garapati, Hannah Gilham, and Jennifer Swanson. Flatiron Health's Catherine Au-Yeung and Tanya Elshahawi contributed to illustration design. A version of the manuscript is currently under consideration as a preprint at medRxiv.org.

Conflict of interest

Authors BA, MW, AB, JK, KK, SN, JG, JR, KH, GH, RL, TB, SW, GA, EE, CB, EF, ME, WS, and AB are employees of Flatiron Health, Inc., which is an independent member of the Roche group, and own stock in Roche.

Publisher's note

References

Adams, G., Alsentzer, E., Ketenci, M., Zucker, J., and Elhadad, N. (2021). What's in a summary? Laying the groundwork for advances in hospital-course summarization. *Proc. Conf.* 2021, 4794–4811. doi:10.18653/v1/2021.naacl-main.382

Adamson, B. J., Cohen, A. B., Cheever, M. A., et al. (2019). "Cancer immunotherapy use and effectiveness in real-world patients living with HIV," Presented at the Abstract Presented at: 17th International Conference on Malignancies in HIV/AIDS. Bethesda, Maryland. October 21-22.

Agrawal, M., Adams, G., Nussbaum, N., et al. (2018). Tifti: A framework for extracting drug intervals from longitudinal clinic notes. arXiv:Preprint posted online Nov 30, 2018

Ambwani, G., Cohen, A., Estévez, M., Singh, N., Adamson, B., Nussbaum, N., et al. (2019). PPM8 A machine learning model for cancer biomarker identification in electronic health records. *Value Health* 22, S334. doi:10.1016/j.jval.2019.04.1631

Ballre, A., Baruah, P., and Amster, G. (2022). *Systems and methods for predicting biomarker status and testing dates*. United States.

Banerjee, I., Bozkurt, S., Caswell-Jin, J., Kurian, A. W., and Rubin, D. L. (2019). Natural language processing approaches to detect the timeline of metastatic recurrence of breast cancer. *JCO Clin. Cancer Inf.* 3, 1–12. doi:10.1200/CCI.19.00034

Benedum, C., Adamson, B., Cohen, A. B., Estevez, M., Sondhi, A., Fidyk, E., et al. (2022). P57 machine learning-accelerated outcomes research: A real-world case study of biomarker-associated overall survival in oncology. *Value Health* 25, S13–S14. doi:10.1016/j.jval.2022.09.069

Benedum, C. M., Sondhi, A., Fidyk, E., Cohen, A. B., Nemeth, S., Adamson, B., et al. (2023). Replication of real-world evidence in oncology using electronic health record data extracted by machine learning. *Cancers (Basel)* 15, 1853. doi:10.3390/cancers15061853

Bera, K., Schalper, K. A., Rimm, D. L., Velcheti, V., Madabhushi, A., et al. (2019). Artificial intelligence in digital pathology - new tools for diagnosis and precision oncology. *Nat. Rev. Clin. Oncol.* 16, 703–715. doi:10.1038/s41571-019-0252-y

Bertsimas, D., and Wiberg, H. (2020). Machine learning in oncology: Methods, applications, and challenges. *JCO Clin. Cancer Inf.* 4, 885–894. doi:10.1200/CCI.20.00072

Bhardwaj, R., Nambiar, A. R., and Dutta, D. (2017). A study of machine learning in healthcare. Presented at the Abstract Presented at: 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC). Turin, Italy. July 4-8.

Birnbaum, B., Nussbaum, N., Seidl-Rathkopf, K., et al. (2020). Model-assisted cohort selection with bias analysis for generating large-scale cohorts from the EHR for oncology research. arXiv:Preprint posted online January 13.

Birnbaum, B. E., and Ambwani, G. (2023). *Generalized biomarker model*. United States.

Birnbaum, B. E., Haimson, J. D., and He, L. D. (2023). *Systems and methods for automatic bias monitoring of cohort models and un-deployment of biased models*. United States.

Birnbaum, B. E., Haimson, J. D., and He, L. D. (2019). *Systems and methods for model-assisted cohort selection*. United States.

Blueprint for trustworthy AI implementation guidance and assurance for healthcare (2022). December 7, update https://www.coalitionforhealthai.org/insights.

Center for Drug Evaluation and Research Center for Biologics Evaluation and Research Oncology Center of Excellence (2022). *Real-world data: Assessing electronic health records and medical claims data to support regulatory decision-making for drug and biological products; draft guidance for industryhttps*. Available at: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-electronic-health-records-and-medical-claims-data-support-regulatory.

Chen, R., Garapati, S., Wu, D., Ko, S., Falk, S., Dierov, D., et al. (2019). Machine learning based predictive model of 5-year survival in multiple myeloma autologous transplant patients. *Blood* 134, 2156. doi:10.1182/blood-2019-129432

Coombs, L., Orlando, A., Wang, X., Shaw, P., Rich, A. S., Lakhtakia, S., et al. (2022). A machine learning framework supporting prospective clinical decisions applied to risk prediction in oncology. *NPJ Digit. Med.* 5, 117. doi:10.1038/s41746-022-00660-3

Datta, S., Bernstam, E. V., and Roberts, K. (2019). A frame semantic overview of NLP-based information extraction for cancer-related EHR notes. *J. Biomed. Inf.* 100, 103301. doi:10.1016/j.jbi.2019.103301

Devlin, J., Chang, M., and Lee, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:Preprint posted online October 11, 2018.

Estévez, M., Benedum, C. M., Jiang, C., Cohen, A. B., Phadke, S., Sarkar, S., et al. (2022). Considerations for the use of machine learning extracted real-world data to support evidence generation: A research-centric evaluation framework. *Cancers (Basel)* 14, 3063. doi:10.3390/cancers14133063

Forsyth, A. W., Barzilay, R., Hughes, K. S., Lui, D., Lorenz, K. A., Enzinger, A., et al. (2018). Machine learning methods to extract documentation of breast cancer symptoms from electronic health records. *J. Pain Symptom Manage* 55, 1492–1499. doi:10.1016/j.jpainsymman.2018.02.016

Gippetti, J., Phadke, S., and Amster, G. (2023). *Systems and methods for extracting dates associated with a patient condition*. United States.

Haimson, J. D., Baxi, S., and Meropol, N. (1997). *Prognostic score based on health information*. United States.

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural comput* 9, 1735–1780. doi:10.1162/neco.1997.9.8.1735

Hooley, I., Chen, R., Long, L., Cohen, A., and Adamson, B. (2019). PCN166 optimization of natural language processing-supported comorbidity classification algorithms in electronic health records. *Value Health* 22, S87. doi:10.1016/j.jval.2019.04.290

Jorge, A., Castro, V. M., Barnado, A., Gainer, V., Hong, C., Cai, T., et al. (2019). Identifying lupus patients in electronic health records: Development and validation of machine learning algorithms and application of rule-based algorithms. *Semin. Arthritis Rheum.* 49, 84–90. doi:10.1016/j.semarthrit.2019.01.002

Karimi, Y. H., Blayney, D. W., Kurian, A. W., Shen, J., Yamashita, R., Rubin, D., et al. (2021). 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 Inf.* 5, 469–478. doi:10.1200/CCI.20.00165

Kelly, J., Wang, C., Zhang, J., Das, S., Ren, A., and Warnekar, P. (2022). Automated mapping of real-world oncology laboratory data to LOINC. *AMIA Annu. Symp. Proc.* 2021, 611–620.

Koleck, T. A., Dreisbach, C., Bourne, P. E., Bakken, S., et al. (2019). Natural language processing of symptoms documented in free-text narratives of electronic health records: A systematic review. *J. Am. Med. Inf. Assoc.* 26, 364–379. doi:10.1093/jamia/ocy173

Lakhanpal, S., Hawkins, K., Dunder, S. G., Donahue, K., Richey, M., Liu, E., et al. (2021). An automated EHR-based tool to facilitate patient identification for biomarker-driven trials. *JCO* 39, 1539. doi:10.1200/jco.2021.39.15_suppl.1539

Lipton, Z. C., Elkan, C., and Naryanaswamy, B. (2014). Optimal thresholding of classifiers to maximize F1 measure. *Mach. Learn Knowl. Discov. Databases* 8725, 225–239. doi:10.1007/978-3-662-44851-9_15

Ma, X., Long, L., and Moon, S. (2023). Comparison of population characteristics in real-world clinical oncology databases in the US: Flatiron health, SEER, and NPCR. medRxiv. Preprint posted online June 07.

Maarseveen, T. D., Maurits, M. P., Niemantsverdriet, E., van der Helm-van Mil, A. H. M., Huizinga, T. W. J., and Knevel, R. (2021). 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.* 23 (1), 174. doi:10.1186/s13075-021-02553-4

NICE (2022). *NICE real-world evidence framework*. Available at: https://www.nice.org.uk/corporate/ecd9/chapter/overview.

Norgeot, B., Quer, G., Beaulieu-Jones, B. K., Torkamani, A., Dias, R., Gianfrancesco, M., et al. (2020). Minimum information about clinical artificial intelligence modeling: The MI-CLAIM checklist. *Nat. Med.* 26, 1320–1324. doi:10.1038/s41591-020-1041-y

Padula, W. V., Kreif, N., Vanness, D. J., Adamson, B., Rueda, J. D., Felizzi, F., et al. (2022). Machine learning methods in health economics and outcomes research—the PALISADE checklist: A good practices report of an ISPOR task force. *Value Health* 25, 1063–1080. doi:10.1016/j.jval.2022.03.022

Rich, A., Amster, G., and Adams, G. (2023). *Deep learning architecture for analyzing unstructured data*. United States.

Rich, A., Leybovich, B., and Irvine, B. (2022). *Machine learning model for extracting diagnoses, treatments, and key dates*. United States.

Rich, A. S., Leybovich, B., Estevez, M., Irvine, J., Singh, N., Cohen, A. B., et al. (2021). Extracting non-small cell lung cancer (NSCLC) diagnosis and diagnosis dates from electronic health record (EHR) text using a deep learning algorithm. *J. Clin. Oncol.* 39, 1556. doi:10.1200/jco.2021.39.15_suppl.1556

Shah, P., Kendall, F., Khozin, S., Goosen, R., Hu, J., Laramie, J., et al. (2019). Artificial intelligence and machine learning in clinical development: A translational perspective. *NPJ Digit. Med.* 2, 69. doi:10.1038/s41746-019-0148-3

Shickel, B., Tighe, P. J., Bihorac, A., Deep, E. H. R., et al. (2018). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE J. Biomed. Health Inf.* 22, 1589–1604. doi:10.1109/JBHI.2017.2767063

Shklarski, G., Abernethy, A., and Birnbaum, B. (2020). *Extracting facts from unstructured data*. United States.

Sondhi, A., Benedum, C., Cohen, A. B., Nemeth, S., Bozkurt, S., et al. (2022). RWD112 can ML-extracted variables reproduce real world comparative effectiveness results from expert-abstracted data? A case study in metastatic non-small cell lung cancer treatment. *Value Health* 25, S470. doi:10.1016/j.jval.2022.09.2337

Subbiah, V. (2023). The next generation of evidence-based medicine. *Nat. Med.* 29, 49–58. doi:10.1038/s41591-022-02160-z

Wang, L., Wampfler, J., Dispenzieri, A., Xu, H., Yang, P., Liu, H., et al. (2019). Achievability to extract specific date information for cancer research. *AMIA Annu. Symp. Proc.* 2019, 893–902. Published 2020 Mar 4, 2020.

Waskom, M. L., Tan, K., Wiberg, H., et al. (2023). A hybrid approach to scalable real-world data curation by machine learning and human experts. *medRxiv:Preprint posted online March* 8. doi:10.1101/2023.03.06.23286770

Yang, R., Zhu, D., Howard, L. E., De Hoedt, A., Williams, S. B., Freedland, S. J., et al. (2022). Identification of patients with metastatic prostate cancer with natural language processing and machine learning. *JCO Clin. Cancer Inf.* 6, 2022, e2100071. doi:10.1200/CCI.21.00071

Zeng, Z., Espino, S., Roy, A., Li, X., Khan, S. A., Clare, S. E., et al. (2018). Using natural language processing and machine learning to identify breast cancer local recurrence. *BMC Bioinforma.* 19, 498–x. doi:10.1186/s12859-018-2466-x

Zhao, J., Agrawal, M., Razavi, P., et al. (2021). Directing human attention in event localization for clinical timeline creation. *PMLR* 149, 80–102.