EXPERT REVIEW OF PHARMACOECONOMICS & OUTCOMES RESEARCH
2024, VOL. 24, NO. 4, 467–475
https://doi.org/10.1080/14737167.2024.2322664

Taylor & Francis
Taylor & Francis Group

SPECIAL REPORT

Using natural language processing to analyze unstructured patient-reported outcomes data derived from electronic health records for cancer populations: a systematic review

Jin-Ah Sim[a,b], Xiaolei Huang[c], Madeline R. Horan[a], Justin N. Baker[d] and I-Chan Huang[a]

[a]Department of Epidemiology and Cancer Control, St. Jude Children's Research Hospital, Memphis, TN, USA; [b]Department of AI Convergence, Hallym University, Chuncheon, Republic of Korea; [c]Department of Computer Science, University of Memphis, Memphis, TN, USA; [d]Department of Pediatrics, Stanford University, Stanford, CA, USA

ABSTRACT

Introduction: Patient-reported outcomes (PROs; symptoms, functional status, quality-of-life) expressed in the 'free-text' or 'unstructured' format within clinical notes from electronic health records (EHRs) offer valuable insights beyond biological and clinical data for medical decision-making. However, a comprehensive assessment of utilizing natural language processing (NLP) coupled with machine learning (ML) methods to analyze unstructured PROs and their clinical implementation for individuals affected by cancer remains lacking.

Areas covered: This study aimed to systematically review published studies that used NLP techniques to extract and analyze PROs in clinical narratives from EHRs for cancer populations. We examined the types of NLP (with and without ML) techniques and platforms for data processing, analysis, and clinical applications.

Expert opinion: Utilizing NLP methods offers a valuable approach for processing and analyzing unstructured PROs among cancer patients and survivors. These techniques encompass a broad range of applications, such as extracting or recognizing PROs, categorizing, characterizing, or grouping PROs, predicting or stratifying risk for unfavorable clinical results, and evaluating connections between PROs and adverse clinical outcomes. The employment of NLP techniques is advantageous in converting substantial volumes of unstructured PRO data within EHRs into practical clinical utilities for individuals with cancer.

ARTICLE HISTORY
Received 2 September 2023
Accepted 20 February 2024

KEYWORDS
Cancer; Electronic health records; machine learning; natural language processing; Patient-reported outcomes

1. Introduction

1.1. Cancer patients and patient-reported outcomes

Evaluating patient-reported outcomes (PROs), defined as self-reported health status, including symptoms, functional status, and quality of life, is useful in oncology because PROs offer distinct insights into individuals grappling with cancer and the repercussions of their health status [1,2]. Clinical trials increasingly incorporate PROs as primary or secondary endpoints [3]. Furthermore, the evaluation of PROs aids healthcare professionals in recognizing the unmet physical, emotional, and social needs of cancer patients and survivors, paving the way for tailored clinical or psychosocial interventions [4].

1.2. Assessment of unstructured PROs for cancer patients

Conventionally, clinicians and researchers rely on validated self-reported PRO measures to collect PROs from cancer patients and survivors [2,5]. However, this approach may not directly capture individual differences in PROs because the PRO content relies on a fixed number of survey items with pre-specified PRO domains and subdomains. It is also challenging to collect PRO data from every cancer patient or survivor during busy clinical encounters [6], which may increase the physical or psychological burden of both the patients and clinicians [7]. Finding another methodology to collect and evaluate PROs, e.g. unstructured or text-based PROs available in EHRs, becomes crucial [8]. Using unstructured PROs as an alternative resource proves clinically practical and beneficial, as regular patient–clinician interactions routinely collect these data. PRO clinical narratives widely exist in EHRs, including admission documents, daily progress notes, clinic visit documentation, discharge summaries, nursing, psychology, or social work notes [9]. Leveraging natural language processing (NLP) coupled with machine learning (ML) techniques offers an opportunity to explore unstructured PROs through EHRs within the oncology context.

1.3. Current status of NLP applications for PRO assessment in oncology

NLP methods can potentially convert unstructured PROs – such as symptom descriptions summarized in physician

CONTACT I-Chan Huang i-chan.huang@stjude.org Department of Epidemiology & Cancer Control, St. Jude Children's Research Hospital, 262 Danny Thomas Place, Memphis, TN 38105, USA

Supplemental data for this article can be accessed online at https://doi.org/10.1080/14737167.2024.2322664

Article highlights

- Unstructured PROs are often passively collected during the patient–clinician conversation and documented as a part of routine clinical care, and a significant number of unstructured PROs have been available in EHRs.
- Due to the challenge of conducting PRO surveys from busy clinical settings, leveraging free-text PROs documented in EHRs and applying NLP for analyzing PROs to improve the cancer decision-making process is clinically relevant.
- Applying NLP methods can greatly enhance the efficiency and precision of examining unstructured PROs data in EHRs for cancer individuals, which will bolster more effective clinical applications in the field of oncology.
- While still in its early stages, the implementation of large language models as a new technique holds the potential to enhance the examination of unstructured PROs and their utilization in oncology.

notes – into a measurable structure supporting clinical research and practice. This process involves categorizing or foreseeing health data through information extraction, semantic representation, and outcome prediction [10]. Utilizing NLP in PRO analysis has the potential to reduce the labor-intensive task of manually reviewing and extracting text-based PRO information from EHRs [11,12]. Feature extraction approaches for extracting measurable properties for PROs from unstructured raw text include non-neural methods (e.g. bag-of-words [BoW]) [13], and the neural ML techniques (e.g. word2vec [W2V] and pre-trained large language models [LLM]) [14,15]. Rule-based extraction approaches analyze PROs within EHRs by employing a set of pre-defined rules, such as simple keyword matching, linguistic patterns (e.g. regular expressions), and dictionary lookups for PRO extraction. For instance, in detecting cancer-related symptoms, rules include recognizing specific keywords or phrases and matching linguistic patterns associated with the symptoms. The rule-based extraction also allows for calculating pain intensity scores for cancer patients by the pre-defined criteria and counting the extracted patterns [16–19]. The advanced techniques show the advantage of effectively managing substantial volumes of unstructured PROs generated by individuals coping with cancer [20].

Several systematic or scoping review studies have reported using NLP to automate extracting clinical/biological information stored in EHRs for cancer populations [21–23]. However, very few studies have focused on NLP for cancer case identification, assessing the cancer staging, or summarizing cancer phenotype. Only one review study has reported the applications of NLP methods for symptoms from EHRs, with a focus on the traditional NLP coupled with non-neural ML techniques (e.g. rule-based ML, Naïve Bayes) [9]. In contrast, our review study includes cutting-edge NLP, typically transformer-based large language models, including Bidirectional Encoder Representations from Transformers (BERT) or Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA). These models excel beyond traditional NLP techniques, which rely on statistical methods (e.g. n-gram, TF-IDF), showcasing superior capabilities to comprehend the context of words within documents.

1.4. Objective

The objective of this investigation is to perform a systematic review of previously published studies utilizing NLP methods to extract and assess unstructured PROs available within the clinical narratives present from EHRs for cancer patients or survivors. We especially identified the specific NLP methods and platforms for data processing, analysis, and subsequent clinical implementation.

2. Methods

2.1. Data retrieval

We searched studies published in English between 1 January 2010 and 31 December 2022, through PubMed, Scopus, and Web of Science databases. Following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines, we identified 1,346 studies from PubMed, 1,777 from Scopus, and 1,758 from Web of Science. We report the search strategies in Supplementary Table S1.

2.2. Article selection

Of the 4,881 studies identified, 3,190 non-duplicate studies were retained before screening the titles and abstracts. We excluded 2,993 studies that did not describe cancer patients or survivors as keywords in the title or abstract. Candidate studies were considered for inclusion if they 1) focused on free-text or unstructured PRO data in clinical notes available from EHRs, 2) used NLP techniques with or without ML algorithms to extract or analyze unstructured PROs, and 3) included cancer patients or survivors. We excluded studies that 1) did not apply NLP techniques, 2) were non-empirical studies (e.g. case reports, commentary, review), 3) were non-EHR-based studies (e.g. patient-authored data collected from social media or online community), 4) were survey-based studies containing quantitative PROs, and 5) focused on non-cancer diseases. Based on these criteria, the first author (JAS) and the senior author (ICH) independently reviewed the abstracts of all 197 studies retrieved from the literature search and retained 30 studies. Subsequently, the same two authors reviewed the full-text articles and kept 22 studies for data extraction (Supplementary Figure S1).

2.3. Data extraction and summary

The first author (JAS) manually extracted and summarized data from 22 selected studies, and the senior author (ICH) confirmed the accuracy of the information extracted. For each selected study, the information related to cancer type, study purpose, the number of narratives, specific domains of PROs, outcomes of interest, NLP systems or toolkits, vocabularies or dictionaries, and performance indicators (e.g. precision, recall, and F1) used to evaluate NLP models were summarized. Using a Sankey diagram plot (with R Studio software), we conducted a synthetic analysis to delineate the association between NLP methods and applications for the studies. We assessed the risk of bias (ROB) issues based on a method proposed in a previous publication [24] to evaluate publication bias of 22 studies included in our review, recognizing no reporting ROB guidelines or tools are available for clinical NLP studies.

3. Results

3.1. Study characteristics

Characteristics of the 22 selected studies are in Supplementary Table S2. These studies included diverse numbers of participants, ranging from 37 to 6,595, and different cancer types, including breast ($n = 6$) [18,25–29], colorectal ($n = 6$) [25–27,30–32], prostate ($n = 3$) [16,19,33], cervical ($n = 2$) [34,35], any types ($n = 6$) [13,15,17,36–38], and other individual types (bone metastases, thoracic cancer) ($n = 2$) [29,39]. The number of documents for NLP applications ranged from 445 to 1,554,736. The sources of datasets include clinical notes from EHRs (81.8%), discharge summaries from EHRs (9.5%), telephone call notes (9.5%), and others (note, in some studies, the datasets were derived from various sources). Only three studies mentioned that their datasets are available for public use [15,30,39]. Unit of the free-text PRO documents included keywords/phrases, sentences, paragraphs, or entire medical notes/documents. For the ROB assessment, each of 22 studies was rated as 'yes' (having low ROB), 'no' (having high ROB), 'unspecified,' and 'not applicable,' respectively (see Supplementary Figure S2). Briefly, most of the studies lack evidence in reporting the annotation process and external validation.

3.2. PRO information extraction and analysis

Supplementary Table S3 presents the domains of unstructured PROs extracted through NLP techniques. The contents of PRO domains included general symptoms ($n = 17$) [13,15,17,18,26–29,31–39], psychological symptoms/functioning ($n = 10$) [13,15,25–27,29,30,32,34,37], digestive/gastrointestinal symptoms ($n = 9$) [13,15–17,26,27,31,37,38], respiratory symptoms ($n = 4$) [13,15,28,29], physical functioning ($n = 4$) [25,27,29,37], urinary symptoms ($n = 4$) [16,19,35,40], head/neck symptoms ($n = 3$) [15,29,33], cardiovascular symptoms ($n = 2$) [15,28], neurocognitive symptoms ($n = 2$) [26,27], social functioning ($n = 2$) [13,37], skin symptoms ($n = 2$) [15,38], musculoskeletal symptoms ($n = 1$) [28], and sexual/reproductive symptoms ($n = 1$) [19].

Figure 1 shows the terminology systems (i.e. vocabulary, dictionary) used by the included studies to capture the concept of PROs, inclusive of the International Classification of Diseases (ICD) codes ($n = 11$) [17–19,25,29,30,32,34–37], the Unified Medical Language System (UMLS) ($n = 6$) [16,26,27,31,33,39], and preexisting dictionaries ($n = 4$) [27,31,39,40].

3.3. Purpose of NLP applications for analyzing PRO data

The approaches or sub-tasks NLP technologies to solve in PRO studies are summarized in Table 1. Over 50% of the included studies used PRO information as a primary outcome of interest, and approximately 30% of the included studies considered PROs as independent risk factors from clinical factors. NLP techniques have been used by all (100%) studies to extract or annotate free-text PROs, followed by using NLP with or without ML to classify, phenotype, or cluster PROs (68.2%), using PRO data to predict the risk of adverse events (e.g. early onset of cancer, type 2 diabetes) (31.8%), stratifying the risk of adverse health events for distinctive patient subgroups (22.7%), and investigating the associations between PROs and clinical outcomes (27.3%).

3.4. NLP/ML techniques

Table 2 summarizes the 3-step NLP/ML applications to analyze the unstructured PRO data. The initial step involves preparing (or preprocessing) unstructured PRO data within medical narratives, such as condensing vocabulary size (e.g. tokenization [19,20,26,29,30], lemmatization, stemming from the corpus [28]) and eliminating disruptions (e.g. punctuation and stopword removal) aiming to capture the underlying linguistic information. However, not all studies have applied preprocessing methods, such as stopwords removal, which is to preserve linguistic and semantic structures. The second step involves feature extraction or representation. Tasks include 1) rule-based extraction methods [13,15,16,18,28,29,31–36,39,40], 2) feature extraction methods inferring whether

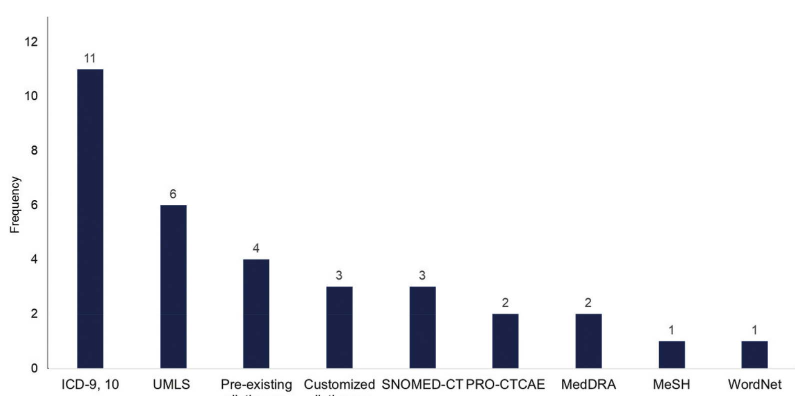

Figure 1. Frequency of vocabulary systems or dictionaries used in 22 studies.

Abbreviations: ICD, International Classification of Diseases; UMLS, Unified Medical Language System; SNOMED-CT, Systematized Nomenclature of Medicine Clinical Terms; PRO-CTCAE, Patient-Reported Outcomes version of the Common Terminology Criteria for Adverse Events; MedDRA, Medical Dictionary for Regulatory Activities; MeSH, Medical Subject Headings.

Table 1. Purpose of applying NLP/ML techniques for processing, annotating, and analyzing unstructured PROs included in the 22 studies*.

Purpose of NLP/ML application	The specific task of NLP/ML application	Task description	N	%
Risk assessment			22	100.0
	PRO content detection, identification, extraction	Detect or identify PRO keywords or terminologies from free text	22	100.0
	PRO annotation	Perform semi-automated or manual labels for PROs in free text	15	68.2
	PRO affirmation/negation	Declare whether symptoms or symptom-related outcomes exist or equivalent expression or negative statement for having symptoms	8	36.4
Classification/phenotyping/ clustering			15	68.2
	PRO classification	Assign or classify extracted PROs into specific categories	13	59.1
	PRO phenotyping	Indicate specific characteristics of single or multiple PRO features	2	9.1
	PRO clustering	Identify two or more PROs that are related to each other or co-occur	3	13.6
NLP/ML pipelines			5	22.7
	Development of NLP/ML pipelines	Develop new NLP/ML pipelines or build NLP software	5	22.7
	Evaluation/validation of NLP/ML pipelines	Evaluate and validate the performances of the NLP system/pipeline	5	22.7
Risk prediction or stratification for clinical outcomes			13	59.1
	Risk prediction	Predict the risk of outcomes using extracted PROs based on unstructured narratives	7	31.8
	Risk stratification	Identify the right level of care and services for distinctive subgroups of patients.	5	22.7
Investigation of associations between PROs and clinical outcomes			6	27.3
	Semantic associations between PROs and clinical outcomes	Detect semantic associations or relationships between unstructured PROs and clinical outcomes	6	27.3

*Some studies may include multiple purposes and NLP/ML tasks.

a named entity of PRO is present or absent (e.g. detecting PRO affirmation/negation [15,16–19,25,26,29,31,34,38,40]), and 3) techniques to address context-free issues that extract data features without considering the surrounding linguistic context, including bag-of-word techniques (e.g. Term Frequency – Inverse Document Frequency [TF-IDF] [13,16,26,36]), context-dependent issues that encode contextual relationships within the text into feature vectors, and Named Entity Recognition [NER] [31,34,35,36,37,39] that involves the extraction of entities and can be vectorized as features, such as the occurrence of specific cancer symptoms. The main purpose of this step is to enrich the data representations as model input. The third step is to develop and deploy NLP models for data analysis. Those models will extract features from unstructured PROs and harness those features for outcome inferences. This step primarily builds NLP/non-neural ML-based classifiers (e.g. Support Vector Machine [SVM], Random Forest [RF] classifier) or NLP/neural network models (e.g. Artificial Neural Network [ANN]) to investigate relationships between PROs and clinical outcomes. The fusion of NLP models with ML techniques has demonstrated their effectiveness in managing substantial amounts of unstructured PRO data in EHRs, enabling broad applications in predicting or classifying health-related outcomes [41].

Overall, 19 studies deployed NLP techniques for preprocessing unstructured PRO data [13,15–19,25,26,28,31–40], 21 studies used NLP techniques for feature extraction and representations [13,15–19,25–29,31–40], 16 studies presented non-neural NLP/ML approaches [13,16–18,25–27,30,32–37,39,40], and 3 studies used neural NLP/ML methods [15,32,37]. Specifically, the most used NLP technique to preprocess unstructured PROs was annotation (e.g. PRO term identification, entity recognition, or relation extraction with different entities mentioned in the text). One of the common techniques for feature extraction and representations involves binarizing affirmation/negation occurrence from text by inferring whether a named entity of PRO is present or absent using the rule-based NLP and named entity recognition (NER). The specific types of named entities may vary based on the objectives of the NLP application in PRO research (e.g. terms of PROs, PRO severity levels, recognition of durations, affected body parts, relation with other clinical outcomes). The most frequent NLP/non-neural ML methods were SVM and conditional random field, whereas ANN and Transformer-based models (e.g. BERT) are common NLP/neural ML approaches in the studies. Figure 2 visualizes the synthetic analysis that applies different NLP/ML techniques and specific sub-tasks to process and analyze unstructured PRO data for other clinical purposes. It summarizes popular techniques with task-driven applications.

Among the 22 studies, 72.7% of them reported the use of different performance metrics (e.g. precision, recall, accuracy, F1) for NLP evaluations [13,15–19,25,26,31–38,40] (Supplementary Table S2). Studies comparing the performance of different NLP/ML pipelines with the same training and test datasets found a superior performance of NLP/neural ML to NLP/non-neural ML based on the accuracy or F1 metrics [32,40].

4. Discussion

Cancer patients and survivors often discuss PRO issues related to treatment, side effects, survivorship, and palliative care-related issues with their healthcare providers. Significant amounts of unstructured PRO data are collected in various clinical narratives of EHRs, while those unstructured data are rarely systematically analyzed as the main domains of interest in oncology [15]. The recent development of novel NLP

Table 2. The 3-step NLP/ML applications and techniques used in the 22 studies*.

Steps and techniques	Description	n	%
Step 1: Preprocessing		**19**	**86.4**
Annotation	The process of assigning labels to indicate specific attributes of PRO elements, typically to train or evaluate ML models in unstructured clinical text.	16	72.7
Entity linking	The process of establishing a translation between terms in different medical coding systems, ontologies, or terminologies (e.g. mapping ICD-10 codes to SNOMED CT concepts) to enable understanding of information across diverse medical-related data.	9	40.9
Part-of-speech (POS) tagging	The task of tagging or labeling a word in a text with its part of speech	5	22.7
Remove stop-words	The process of excluding common words ('a,' 'the,' 'in,' 'are,' etc.) which are irrelevant for data analysis from the text data to focus on the more significant words that carry more semantic meaning.	5	22.7
Normalization	The process of transforming text data into a standard, consistent format	3	13.6
Lemmatization/stemming	The process of grouping related words by reducing words to their base or root form (e.g. lemmatization, reducing words to their base or dictionary forms; stemming, removing prefixes or suffixes)	1	4.5
Step 2: Feature extraction and representations		**21**	**95.5**
Rule-based NLP	The process of creating explicit rules based on linguistic patterns and medical knowledge to identify and extract mentions of PRO information within EHRs.	14	63.6
Detecting affirmation/negation	The process of classifying whether PROs mentioned in EHRs are affirmed (present) or negated (absent) within the clinical narratives.	12	54.5
Named entity recognition (NER)	The process of identifying or extracting PRO entities or mentions within EHRs and other clinical texts.	6	27.3
Word embedding (e.g. Word2vec, FastText, GloVe)	The unsupervised process of capturing semantic relationships between words in text corpus and mapping them into numerical vectors.	6	27.3
N-gram	It refers to a series of adjacent letters and tokens (e.g. words and symbols).	4	18.2
Topic modeling (e.g. Latent Dirichlet Allocation [LDA])	The process of identifying latent topics or themes within a collection of text data related to PROs.	1	4.5
Step 3: Model development		**17**	**77.3**
NLP/Non-neural ML	It refers to traditional, or statistical approaches that do not involve neural network.	**16**	**72.7**
Logistic regression classifier	A supervised machine learning model by logistic regression that can categorize text documents into predefined binary or multiple classes.	6	27.3
Conditional random fields (CRF)	A statistical modeling method often applied in sequence labeling for PRO data (e.g. descriptions of patients' symptoms, or quality of life), such as identifying specific attributes within clinical narratives.	4	18.2
Rule-based classifier (e.g., regular expression, association rule mining)	A type of model that makes its decisions based on a set of human defined rules, such as scoring threshold to assist if-then decisions.	4	18.2
Support vector machine (SVM)	A supervised machine learning model by SVM that finds the best hyperplane in vector spaces to categorize data entries.	3	13.6
Naïve Bayesian classifier	A probabilistic classification algorithm based on Bayes' theorem, which assumes independence between features.	3	13.6
Random forest (RF) classifier	An ensemble learning algorithm made up of decision trees for classification and regression tasks.	3	13.6
Boosting (e.g. Light Gradient Boosting Machine [LightGBM], eXtreme Gradient Boosting [XGBoost])	An ensemble learning algorithm to combine weak learners (e.g. decision trees) to create a strong learner.	2	9.1
K-means clustering	The process of grouping data entries K sets based on their similarities.	2	9.1
Decision tree (DT) classifier	A non-parametric supervised learning algorithm for classification and regression tasks of NLP.	1	4.5
NLP/Neural ML		**3**	**13.6**
Artificial neural network (ANN) (e.g. Feed forward network [FFN])	A class of ML models inspired by the structure and function of the human brain. It can be used for text classifications, NER, word embeddings, text clustering, etc.	2	9.1
Transformer-based language model (e.g. Bidirectional Encoder Representations from Transformers [BERT], Efficiently Learning an Encoder that Classifies Token Replacements Accurately [ELECTRA], eXtreme Learning with Large-scale Pre-trained Networks [XLNet])	A type of deep learning architecture, which uses Transformer modules equipped with the multi-head attention mechanism. The model has a strong capability that learns meaning and context by tracking relationships in sequential data.	1	4.5

*Some studies may include 2 steps only.

techniques has facilitated the processing of unstructured PRO information (e.g. annotation) for further clinical applications (e.g. the prediction of adverse outcomes) if traditional, quantitative PRO data are not available [10–12]. This systematic review study enhances the existing body of literature by summarizing specific NLP approaches employed to tackle diverse research inquiries related to PRO investigation in oncology.

In contrast to the previous review study [9] summarizing NLP methods to analyze symptoms in EHRs of all types of diseases, our review study includes NLP techniques to process a broader category of unstructured PROs, including patients' symptoms, functioning, and health-related quality of life restricted to the cancer population. In addition, the prior review article focused on traditional NLP techniques, rather than cutting-edge NLP technologies for clinical applications. A prior review study [9] collected the articles published between 1999 and 2017, examined the NLP applications to process or analyze symptoms in medical notes from EHRs, and does not target at the cancer. In contrast, our review study

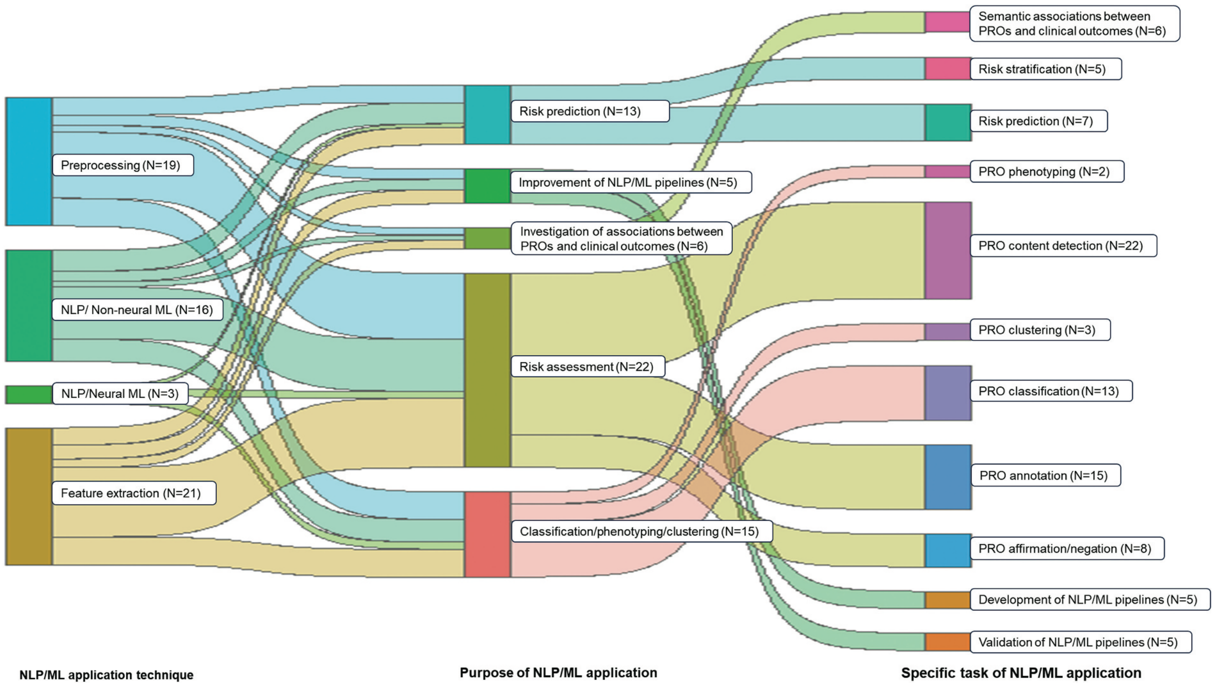

Figure 2. Synthetic analysis for the use of NLP/ML techniques and the purpose and task of NLP/ML applications*.

*This Sankey plot displays the relationships among NLP/ML application techniques, the purpose of NLP/ML applications, and the specific task of NLP/ML applications from 22 studies. We summarized data from 22 studies and grouped their combinations of NLP/ML application techniques (4 categories), the purpose of NLP/ML applications (5 categories), and the specific task of NLP/ML applications (11 categories). Therefore, there are up to 220 ($4 \times 5 * 11$) different relationships among 22 studies.

focuses on the articles published between 2000 and 2022 and covers research involving cancer patients and survivors. In addition, the previous review study did not focus on 'EHRs' as the context. Therefore, the previous review sourced 27 studies and our review included 22 studies, with an overlap of 2 studies. Extending the scope of NLP applications to the analysis of unstructured PROs from EHRs for the cancer population is clinically impactful.

4.1. NLP/ML techniques

In our systematic review, more than 95% of the studies used feature extraction and representation methods to transform unstructured PROs into structured values and applied NLP/non-neural ML (77%) or NLP/neural ML (14%) techniques to predict or classify clinical outcomes. The traditional feature extraction and representation methods (e.g. TF-IDF, NER, n-gram, rule-based) require domain knowledge to design complex feature engineering with limited flexibility [42]. In the context of generic classifiers (e.g. SVM, RNN), the goal is to categorize PROs into different classes (e.g. classifying paragraphs containing PROs, indicating the presence of any of PROs, and annotating more specific spans of text with UMLS concept unique identifiers from the text documents). Several studies in our review reported superior performance by neural versus non-neural models in classifying unstructured PROs or predicting clinical outcomes with or without clinical data from EHRs [32,40]. Redd et al. [32] demonstrated an increase in the accuracy of their NLP algorithm ranging from 0.87 to 0.98 when replacing non-neural models (e.g. Logistic Regression, SVM, and Random Forest) by a deep neural network to classify colorectal cancer cases based on the extracted topic features. Recently, state-of-the-art neural ML approaches, including LLMs, have been increasingly used in oncology to identify cancer staging [43,44] or radiological features [45] of cancer patients, but they have not been popularly used to identify PROs. Our review study found that only one study reported the use of BERT to detect cancer-related symptoms from clinical narratives based on the PRO-CTCAE rubric and compared that to the performances of other LLMs (e.g. Distilled BERT [DistilBERT], Robustly Optimized BERT pretraining approach [RoBERTa], ELECTRA, eXtreme Language understanding NETwork [XLNet]) to identify PROs (e.g. diarrhea, dizziness, and nausea) in the external validation set of MIMIC-III (F1 values of 0.97) [15].

4.2. Vocabulary/Dictionaries applied in NLP studies

Approximately 96% of the included studies used medical vocabularies or dictionaries per standard ontologies, medical terminologies/nomenclatures, or created their own rules to process unstructured PRO data. To achieve semantic interoperability [46], most studies used a medical vocabulary or dictionary system to map PRO words or terminologies [15–19,26–34,36–39]. These vocabulary or dictionary systems help identify the meaning of the words and terms from free-text PROs related to cancer [47,48]. In our review, the ICD codes, UMLS, and Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) were the most widely used vocabulary or dictionary systems [16–19,25–27,29–39]. However, to account for cancer-specific contents, several studies customized their dictionaries and rules by incorporating additional PRO-related vocabularies and terminologies to complement the extant clinical dictionary or vocabulary systems [16,18,19,27,29,33,34,38–40]. Utilizing or developing domain-specific ontologies or dictionaries for annotating unstructured PROs related to oncology can enhance the precision of PRO classification, phenotyping, and subsequent clinical utility [19].

4.3. Limitations

There are two major constraints of this study. First, the 22 studies under our review present specific types of adult-onset cancers (e.g. breast, colorectal, prostate). Therefore, the findings may not be generalizable to other adult cancer diagnoses (e.g. lung, stomach) or pediatric, adolescent, and young adult cancers. Second, we did not evaluate the quality of the selected studies, because the standards or guidelines for evaluating NLP applications in unstructured PROs are understudied. Additional efforts will develop standards or guidelines for assessing NLP applications in analyzing unstructured PROs and their clinical use for cancer patients.

4.4. Conclusions

This systematic review analyzed and summarized the NLP techniques from 22 publications on applying NLP techniques to process or analyze PRO information from clinical narratives in EHRs for cancer populations. Instead of relying on the traditional survey approaches, NLP techniques can automatically process a significant amount of unstructured PRO across diverse data structures and types of cancer (e.g. PRO extraction, classification, risk prediction or stratification for adverse events, and association identification between PROs and adverse clinical outcomes). The diverse patterns of NLP applications suggest that using NLP pipelines to extract and analyze PRO data with integrating findings into EHRs may facilitate the integration of PROs into clinical care and improve the clinical decision-making process and healthcare quality for cancer populations.

4.5. Expert opinion

The findings of this review study provide significant clinical and research implications. To overcome the challenges of collecting PRO data in the busy oncology setting, future efforts may replace PRO surveys with unstructured PROs routinely collected from the daily clinical practice for research and clinical applications. Applying NLP techniques is essential for using unstructured PRO data effectively in oncology. Readers may refer to Table 1 and Table 2 for specific NLP (with or without ML) techniques and their corresponding PRO application tasks, and Figure 2 for an integrated information of Table 1 and Table 2. This figure helps researchers and clinicians to decide the use of specific NLP technologies to solve the purposes or tasks of PRO analysis through EHR architectures. Despite the continuous advancement of NLP methods, effectively applying these techniques demands establishing cohesive systems that seamlessly integrate NLP algorithms into the EHR systems, which is vital for streamlining clinical analysis and interpretation. Given the complexity of processing and analyzing unstructured PROs from fragmented or heterogeneous words or sentences describing a patient's health status or symptoms [19,49], adopting or establishing a high-quality standard for obtaining, processing, and analyzing unstructured PROs is crucial. Adopting the Common Data Model to enhance the quality of PRO data storage coupled with appropriate data mining processing to conduct data analysis should be conducted [50–52].

A few studies ($n = 4$) [15,32,37,39] used advanced NLP to unlock complex unstructured PRO data for cancer populations. Advanced approaches such as Transformer-based models (e.g. BERT and Generative Pre-trained Transformers [GPT]) are increasingly being employed in the realm of oncology research, which facilitates extracting oncological outcomes and providing answers to patients' oncology-related queries [53,54]. However, those LLMs have not been widely adopted in the studies under our review, though one study included in our review building a transformer-based model (e.g. generic BERT, RoBERTa, ELECTRA) has achieved success in using embedding techniques to learn representative features of free-text PROs [15]. A recent study by Lu et al. [20] using semi-structured interview data (non-EHR-based) found that the BERT performed superior to the other methods in identifying different attributes of pain interference and fatigue experienced by pediatric cancer survivors. While the BERT model shows impressive gains in performance (e.g. higher precision, recall, and F1 measure) over many traditional NLP methods, we need to be cautious of several considerations. Firstly, LLMs require significant computational resources, typically high-performance GPUs for efficient text processing. Second, the model size may cause another barrier. Model distillation techniques such as DistilBERT may resolve this challenge [15]. Finally, pretraining efforts still rely on the corpora of generic domains (e.g. news or Wikipedia articles) that may not be generalizable to cancer populations [55]. The use of domain-specific transformer-based models (e.g. Bio-BERT, Cancer-BERT, Med-BERT) may improve the accuracy in analyzing unstructured PROs for cancer populations [44,55–57]. Utilizing and analyzing the unstructured PROs over time through NLP may help evaluate the longitudinal patterns of PROs, which relies on the successful implementation task of cross-sectional PRO patterns for each patient. This approach will offer the opportunity to identify the worsening PROs as early indications of progressive clinical outcomes for cancer patients and survivors, leading to the creation of an individualized plan for cancer screening, treatment, or lifestyle intervention. Unfortunately, most of the studies in our review merely used one-time PRO data derived from EHRs, except for a few studies that considered multiple time points from EHRs [17,27,32,33,39]. Given the worsened symptoms and declining quality-of-life among cancer individuals over time [58], future effort is warranted to include unstructured PRO data throughout the cancer journey with appropriate longitudinal NLP techniques to improve clinical management.

Finally, there is a need to assess the validity or clinical relevance of NLP programs designed to examine unstructured PROs versus quantitative PROs collected from a patient's self-reported survey or questionnaire with structured features for cancer individuals. Several studies in our review have noted the value of using unstructured PROs in predicting or correlating various clinical outcomes (e.g. cancer stage, early onset of cancer, readmission). These findings suggest that extracted unstructured PROs from EHRs can be a surrogate for standard PRO surveys. When using AI-driven methods for extracting, annotating, and analyzing unstructured PROs, it is crucial to establish a guideline or standard to facilitate the choice of appropriate NLP methods, data sources, and evaluation metrics.

Funding

The research reported in this manuscript was supported by the U.S. National Cancer Institute R01CA238368 (Huang/Baker) and the National Science Foundation IIS-2245920 (Huang). The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

Declarations of interest

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

Author contributions

Conceptualization: Jin-ah Sim, I-Chan Huang; Data curation: Jin-ah Sim; Funding acquisition: I-Chan Huang, Xiaolei Huang; Methodology: Jin-ah Sim, Xiaolei Huang, I-Chan Huang; Project administration: I-Chan Huang; Resources: I-Chan Huang; Supervision: I-Chan Huang; Visualization: Jin-ah Sim; Writing – original draft preparation: Jin-ah Sim, I-Chan Huang; Writing – review and editing: Xiaolei Huang, Madeline R. Horan, Justin. N. Baker, I-Chan Huang; All authors have read and agreed to the submitted version of the manuscript.

Reviewer disclosures

Peer reviewers on this manuscript have no relevant financial or other relationships to disclose.

References

Papers of special note have been highlighted as either of interest (•) or of considerable interest (••) to readers.

1. Silveira A, Sequeira T, Gonçalves J, et al. Patient reported outcomes in oncology: changing perspectives—a systematic review. Health Qual Life Outcomes. 2022;20(1):82. doi:10.1186/s12955-022-01987-x
2. Minvielle E, di Palma M, Mir O, et al. The use of patient-reported outcomes (PROs) in cancer care: a realistic strategy. Ann Oncol. 2022;33(4):357–359. doi:10.1016/j.annonc.2021.12.010
3. Mercieca-Bebber R, King MT, Calvert MJ, et al. The importance of patient-reported outcomes in clinical trials and strategies for future optimization. Patient Relat Outcome Meas. 2018;9:353–367. doi:10.2147/PROM.S156279
4. Basch E, Deal AM, Dueck AC, et al. Overall survival results of a trial assessing patient-reported outcomes for symptom monitoring during routine cancer treatment. JAMA. 2017;318(2):197–198. doi: 10.1001/jama.2017.7156
5. Tian L, Cao X, Feng X, et al. Evaluation of psychometric properties of needs assessment tools in cancer patients: a systematic literature review. PloS One. 2019;14(1):e0210242. doi:10.1371/journal.pone.0210242
6. Foster A, Croot L, Brazier J, et al. The facilitators and barriers to implementing patient reported outcome measures in organisations delivering health related services: a systematic review of reviews. J Patient Rep Outcomes. 2018;2(1):46. doi:10.1186/s41687-018-0072-3
7. Cheung YT, Chan A, Charalambous A, et al. The use of patient-reported outcomes in routine cancer care: preliminary insights from a multinational scoping survey of oncology practitioners. Support Care Cancer. 2022;30(2):1427–1439. doi: 10.1007/s00520-021-06545-7
8. Alzu'bi AA, Watzlaf VJM, Sheridan P. Electronic health record (EHR) abstraction. Perspect Health Inf Manag. 2021;18(Spring):1g.
9. Koleck TA, Dreisbach C, Bourne PE, et al. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. J Am Med Inform Assoc. 2019;26(4):364–379.
   - •• **This review study summarizes the use of NLP methods to analyze symptoms in EHRs for any type of disease.**
10. Kong HJ. Managing unstructured big data in healthcare system. Healthc Inform Res. 2019;25(1):1–2. doi:10.4258/hir.2019.25.1.1
11. Juhn Y, Liu H. Artificial intelligence approaches using natural language processing to advance EHR-based clinical research. J Allergy Clin Immunol. 2020;145(2):463–469. doi:10.1016/j.jaci.2019.12.897
12. Schmajuk G, Yazdany J. Leveraging the electronic health record to improve quality and safety in rheumatology. Rheumatol Int. 2017;37(10):1603–1610. doi:10.1007/s00296-017-3804-4
13. Masukawa K, Aoyama M, Yokota S, et al. Machine learning models to detect social distress, spiritual pain, and severe physical psychological symptoms in terminally ill patients with cancer from unstructured text data in electronic medical records. Palliat Med. 2022;36(8):1207–1216. doi: 10.1177/02692163221105595
14. Shuai Z, Xiaolin D, Jing Y, et al. Comparison of different feature extraction methods for applicable automated ICD coding. BMC Med Inform Decis Mak. 2022;22(1):11. doi: 10.1186/s12911-022-01753-5
15. Lindvall C, Deng CY, Agaronnik ND, et al. Deep learning for cancer symptoms monitoring on the basis of electronic health record unstructured clinical notes. JCO Clin Cancer Inform. 2022;6(6):e2100136. doi: 10.1200/CCI.21.00136
   - •• **This study built a transformer-based model and shown a success in using embedding techniques to learn representative features of free-text PROs.**
16. Banerjee I, Li K, Seneviratne M, et al. Weakly supervised natural language processing for assessing patient-centered outcome following prostate cancer treatment. JAMIA Open. 2019;2(1):150–159. doi: 10.1093/jamiaopen/ooy057
17. DiMartino L, Miano T, Wessell K, et al. Identification of uncontrolled symptoms in cancer patients using natural language processing. J Pain Symptom Manage. 2022;63(4):610–617. doi:10.1016/j.jpainsymman.2021.10.014
18. Forsyth AW, Barzilay R, Hughes KS, et al. Machine learning methods to extract documentation of breast cancer symptoms from electronic health records. J Pain Symptom Manage. 2018;55(6):1492–1499. doi: 10.1016/j.jpainsymman.2018.02.016
19. Hernandez-Boussard T, Kourdis PD, Seto T, et al. Mining electronic health records to extract patient-centered outcomes following prostate cancer treatment. AMIA Annu Symp Proc. 2017;2017:876–882.
20. Lu Z, Sim JA, Wang JX, et al. Natural language processing and machine learning methods to characterize unstructured patient-reported outcomes: validation study. J Med Internet Res. 2021;23(11):e26777. doi: 10.2196/26777
   - • **This study used semi-structured interview data to demonstrate the superior performance of the BERT to the other NLP/ML methods in identifying different attributes of pain interference and fatigue experienced by pediatric cancer survivors.**
21. Wang L, Fu S, Wen A, et al. Assessment of electronic health record for cancer research and patient care through a scoping review of

cancer natural language processing. JCO Clin Cancer Inform. 2022;6(6):e2200006. doi: 10.1200/CCI.22.00006
22. Saha A, Burns L, Kulkarni AM. A scoping review of natural language processing of radiology reports in breast cancer. Front Oncol. 2023;13:1160167. doi:10.3389/fonc.2023.1160167
23. Yim WW, Yetisgen M, Harris WP, et al. Natural language processing in oncology: a review. JAMA Oncol. 2016;2(6):797–804. doi:10.1001/jamaoncol.2016.0213
24. Davidson EM, Poon MTC, Casey A, et al. The reporting quality of natural language processing studies: systematic review of studies of radiology reports. BMC Med Imaging. 2021;21(1):142. doi: 10.1186/s12880-021-00671-8
25. Leis A, Casadevall D, Albanell J, et al. Exploring the association of cancer and depression in electronic health records: combining encoded diagnosis and mining free-text clinical notes. JMIR Cancer. 2022;8(3):e39003. doi: 10.2196/39003
26. Luo X, Gandhi P, Storey S, et al. A computational framework to analyze the associations between symptoms and cancer patient attributes post chemotherapy using EHR data. IEEE J Biomed Health Inform. 2021;25(11):4098–4109. doi:10.1109/JBHI.2021.3117238
27. Luo X, Storey S, Gandhi P, et al. Analyzing the symptoms in colorectal and breast cancer patients with or without type 2 diabetes using EHR data. Health Inform J. 2021;27(1):14604582211000785. doi:10.1177/14604582211000785
28. Oyelade ON, Obiniyi AA, Junaidu SB, et al. Patient symptoms elicitation process for breast cancer medical expert systems: a semantic web and natural language parsing approach. Future Computing Inform J. 2018;3(1):72–81. doi:10.1016/j.fcij.2017.11.003
29. Tamang S, Patel MI, Blayney DW, et al. Detecting unplanned care from clinician notes in electronic health records. J Oncol Pract. 2015;11(3):e313–319. doi: 10.1200/JOP.2014.002741
30. Agaronnik ND, El-Jawahri A, Lindvall C, et al. Exploring the process of cancer care for patients with pre-existing mobility disability. JCO Oncol Pract. 2021;17(1):e53–e61. doi:10.1200/OP.20.00378
31. Becker M, Kasper S, Böckmann B, et al. Natural language processing of German clinical colorectal cancer notes for guideline-based treatment evaluation. Int J Med Inform. 2019;127:141–146. doi:10.1016/j.ijmedinf.2019.04.022
32. Redd DF, Shao Y, Zeng-Treitler Q, et al. Identification of colorectal cancer using structured and free text clinical data. Health Inform J. 2022;28(4):14604582221134406. doi: 10.1177/14604582221134406
- **This study demonstrates an increased accuracy in NLP algorithm when replacing non-neural models by a deep neural network for classifing colorectal cancer cases based on the extracted topic features.**
33. Heintzelman NH, Taylor RJ, Simonsen L, et al. Longitudinal analysis of pain in patients with metastatic prostate cancer using natural language processing of medical record text. J Am Med Inform Assoc. 2013;20(5):898–905. doi: 10.1136/amiajnl-2012-001076
34. Weegar R, Kvist M, Sundström K, et al. Finding cervical cancer symptoms in Swedish clinical text using a machine learning approach and NegEx. AMIA Annu Symp Proc. 2015;2015:1296. doi:10.1016/j.jbi.2014.01.01
35. Hong N, Chang F, Ou Z, et al. Construction of the cervical cancer common terminology for promoting semantic interoperability and utilization of Chinese clinical data. BMC Med Inform Decis Mak. 2021;21(9):309. doi: 10.1186/s12911-021-01672-x
36. Jensen K, Soguero-Ruiz C, Oyvind Mikalsen K, et al. Analysis of free text in electronic health records for identification of cancer patient trajectories. Sci Rep. 2017;7(1):46226. doi: 10.1038/srep46226
37. Sivakumar K, Nithya NS, Revathy O. Phenotype algorithm based big data analytics for cancer diagnose. J Med Syst. 2019;43(8):264. doi:10.1007/s10916-019-1409-z
38. Hong JC, Fairchild AT, Tanksley JP, et al. Natural language processing for abstraction of cancer treatment toxicities: accuracy versus human experts. JAMIA Open. 2020;3(4):513–517. doi:10.1093/jamiaopen/ooaa064
39. Naseri H, Kafi K, Skamene S, et al. Development of a generalizable natural language processing pipeline to extract physician-reported pain from clinical reports: generated using publicly-available datasets and tested on institutional clinical reports for cancer patients with bone metastases. J Biomed Inform. 2021;120:103864. doi: 10.1016/j.jbi.2021.103864
40. Li K, Banerjee I, Magnani CJ, et al. Clinical documentation to predict factors associated with urinary incontinence following prostatectomy for prostate cancer. Res Rep Urol. 2020;12:7–14. doi:10.2147/RRU.S234178
41. Le Glaz A, Haralambous Y, Kim-Dufor DH, et al. Machine learning and natural language processing in mental health: systematic review. J Med Internet Res. 2021;23(5):e15708. doi: 10.2196/15708
42. Gasparetto A, Marcuzzo M, Zangari A, et al. A survey on text classification algorithms: from text to predictions. Information. 2022;13(2):83. doi:10.3390/info13020083
43. Hu D, Zhang H, Li S, et al. Automatic extraction of lung cancer staging information from computed tomography reports: deep learning approach. JMIR Med Inform. 2021;9(7):e27955. doi:10.2196/27955
44. Zhou S, Wang N, Wang L, et al. CancerBERT: a cancer domain-specific language model for extracting breast cancer phenotypes from electronic health records. J Am Med Inform Assoc. 2022;29(7):1208–1216. doi:10.1093/jamia/ocac040
45. Ren J, Chen L, Xu H, et al. A bi-LSTM and multihead attention-based model incorporating radiomics signatures and radiological features for differentiating the main subtypes of lung adenocarcinoma. Quant Imaging Med Surg. 2023;13(7):4245–4256. doi:10.21037/qims-22-848
46. Tayefi M, Ngo P, Chomutare T, et al. Challenges and opportunities beyond structured data in analysis of electronic health records. WIREs Comput Stat. 2021;13(6):e1549. doi: 10.1002/wics.1549
47. Slaughter L, Ruland C, Rotegård AK. Mapping cancer patients' symptoms to UMLS concepts. AMIA Annu Symp Proc. 2005;2005:699–703.
48. Silva MC, Eugénio P, Faria D, et al. Ontologies and knowledge graphs in oncology research. Cancers (Basel). 2022;14(8):1906. doi: 10.3390/cancers14081906
49. Ajami S, Arab-Chadegani R. Barriers to implement Electronic Health Records (EHRs). Mater Sociomed. 2013;25(3):213–215. doi:10.5455/msm.2013.25.213-215
50. Martin-Sanchez FJ, Aguiar-Pulido V, Lopez-Campos GH, et al. Secondary use and analysis of big data collected for patient care. Yearb Med Inform. 2017;26(1):28–37. doi:10.15265/IY-2017-008
51. Landolsi MY, Hlaoua L, Ben Romdhane L. Information extraction from electronic medical documents: state of the art and future research directions. Knowl Inf Syst. 2023;65(2):463–516. doi:10.1007/s10115-022-01779-1
52. Schneeweiss S, Brown JS, Bate A, et al. Choosing among common data models for real-world data analyses fit for making decisions about the effectiveness of medical products. Clin Pharmacol Ther. 2020;107(4):827–833. doi:10.1002/cpt.1577
53. Holmes J, Liu Z, Zhang L, et al. Evaluating large language models on a highly-specialized topic, radiation oncology physics. Front Oncol. 2023;13. doi: 10.3389/fonc.2023.1219326
54. Araki K, Matsumoto N, Togo K, et al. Developing artificial intelligence models for extracting oncologic outcomes from Japanese electronic health records. Adv Ther. 2023;40(3):934–950. doi: 10.1007/s12325-022-02397-7
55. Gu Y, Tinn R, Cheng H, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthc. 2021;3(1):1–23. doi: 10.1145/3458754
56. Rasmy L, Xiang Y, Xie Z, et al. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ Digit Med. 2021;4(1):86. doi:10.1038/s41746-021-00455-y
57. Mitchell JR, Szepietowski P, Howard R, et al. A question-and-answer system to extract data from free-text oncological pathology reports (CancerBERT network): development study. J Med Internet Res. 2022;24(3):e27210. doi: 10.2196/27210
58. Tam S, Zatirka T, Neslund-Dudas C, et al. Real time patient-reported outcome measures in patients with cancer: early experience within an integrated health system. Cancer Med. 2023;12(7):8860–8870. doi: 10.1002/cam4.5635